目　　录

第一章 基本概念

§1.1 统计结构

§1.1.1 统计结构

概率论和数理统计都是研究随机现象统计规律性的数学学科，它们之间联系密切，但也有根本差别：在概率论中研究的出发点是一个概率空间$(\mathscr{X},\mathscr{B},\mathscr{P})$，即已知一个样本空间$\mathscr{X}$、$\mathscr{X}$中某些子集组成的$\sigma$代数$\mathscr{B}$和在可测空间$(\mathscr{X},\mathscr{B})$上定义的一个概率分布$P$，然后研究这个概率空间的各种性质；而在数理统计中研究的是一组受到随机性干扰的数据，再加上人们对这组数据的一些认识(即各种假设)，就形成数理统计研究的出发点，然后对所考虑的问题作出统计推断或预测. 为说清这个出发点，我们先看一个例子.

例 1.1 （测量问题）一个试验者对未知的物理量μ进行测量，为了对μ作出估计，大家知道，他的测量值x会受到各种随机因素的影响，以至于使x可认为是μ加上随机误差ϵ后而得到的，即

$$x=\mu+\epsilon$$

这里的“可加性”是人们对测量数据构成所作的一个假设，经过多次使用经验，说明这个假设是合理的. 另外，由于测量误差ϵ是受到测量仪器、环境温度、光线、视觉、心理等因素的微小变化而引起的综合结果，据中心极限定理，又可认为ϵ是服从均值为0和方差为σ^2的正态分布$N(0,\sigma^2)$. 这是人们对测量数据的另一个假设(认识). 至此，我们对这个测量值问题的认识有如下三点：

1. 测量值x可取任何实数，实数集$\mathbf{R}$组成样本空间；

2. 实数集 $\mathbf{R}$ 上的 Borel 集的全体组成的 σ 代数 $\mathscr{B}_{\mathbf{R}}$；

3. 在可测空间$(\mathbf{R},\mathscr{B}_{\mathbf{R}})$上的一个概率分布族

$$\mathscr{P}_1=\{N(\mu,\sigma^2):(\mu,\sigma)\in\mathbf{R}\times\mathbf{R}^+\}$$

其中 $\mathbf{R}^+$ 是正实数集.

这样三件东西 $\mathbf{R}$，$\mathscr{B}_{\mathbf{R}}$ 和 $\mathscr{P}_1$ 就是我们研究测量问题的出发点. 假如不仅了解 ε 服从正态分布，而且还知其方差为 σ_0^2（比如知道测量仪器的精度），那么分布族就缩小为

$$\mathscr{P}_2=\{N(\mu,\sigma_0^2):\mu\in\mathbf{R}\}$$

这时 $\mathbf{R}$，$\mathscr{B}_{\mathbf{R}}$ 和 $\mathscr{P}_2$ 就成为我们研究这个问题的出发点，假如我们对随机误差 ε 了解甚少，讲不出 ε 的分布是什么类型，只知道它是关于 0 对称的连续分布，那么分布族就扩大为

$$\mathscr{P}_3=\{P:P\ \text{为}\ \mathbf{R}\ \text{上关于}\ \mu\ \text{对称的分布}\}$$

这时，研究这个问题的出发点就是 $\mathbf{R}$，$\mathscr{B}_{\mathbf{R}}$ 和 $\mathscr{P}_3$.

定义 1.1 设$(\mathscr{X},\mathscr{B})$为可测空间，$\mathscr{P}$ 为其上的一个概率分布族，则称三元组$(\mathscr{X},\mathscr{B},\mathscr{P})$为统计结构，或称为统计模型. 假如分布族 $\mathscr{P}$ 仅依赖于某个参数（或参数向量）θ，即

$$\mathscr{P}=\{P_\theta:\theta\in\Theta\}$$

其中 Θ 为参数空间，则称此结构为参数（统计）结构，或称为参数（统计）模型，否则称为非参数（统计）结构或非参数（统计）模型.

在例 1.1 中，$(\mathbf{R},\mathscr{B}_{\mathbf{R}},\mathscr{P}_1)$，$(\mathbf{R},\mathscr{B}_{\mathbf{R}},\mathscr{P}_2)$，$(\mathbf{R},\mathscr{B}_{\mathbf{R}},\mathscr{P}_3)$是三个不同的统计结构. 它们之间的差别可能反映实际背景的差别，也可能反映人们对实际情况认识上的差别，因此这三个结构可能都是合理的，至于选用哪一个结构，这已不是一个理论问题，而是一个实践性很强的问题，人们常凭借经验积累、专业知识和抽象概括等来确定统计结构.

在例 1.1 中，$(\mathbf{R},\mathscr{B}_{\mathbf{R}},\mathscr{P}_1)$和$(\mathbf{R},\mathscr{B}_{\mathbf{R}},\mathscr{P}_2)$是参数结构，而$(\mathbf{R},\mathscr{B}_{\mathbf{R}},\mathscr{P}_3)$是非参数结构. 人们往往希望从参数结构出发来研究问题，因为参数结构含有较多的信息，由此出发，可以获得精度较高的参数估计，但这样做要意识到是有风险的，因为当参数结构不真时，那推断结果可能离实际更远了. 若选用非参数结构，所冒风险就要小得多，因为非参数结构所含的信息较少，适应面广，但精度一般不会很高，以后会看到，在

这两类结构下所用的统计推断方法有很大差别,以至于在今天已形成统计中的参数方法与非参数方法两类.

§1.1.2 乘积结构与重复抽样结构

由简单的统计结构可以派生出一些比较复杂的统计结构.

定义 1.2 设$(\mathscr{X},\mathscr{B},\mathscr{P})$和$(\mathscr{X}',\mathscr{B}',\mathscr{P}')$是两个统计结构,则称$(\mathscr{X}\times\mathscr{X}',\mathscr{B}\otimes\mathscr{B}',\mathscr{P}\otimes\mathscr{P}')$为两者的乘积结构,并记为$(\mathscr{X},\mathscr{B},\mathscr{P})\otimes(\mathscr{X}',\mathscr{B}',\mathscr{P}')$,其中

$$\mathscr{P}\otimes\mathscr{P}'=\{P\otimes P':P\in\mathscr{P},P'\in\mathscr{P}'\}$$

类似地可以给出多于两个统计结构的乘积结构.特别,n个相同统计结构$(\mathscr{X},\mathscr{B},\mathscr{P})$的乘积结构称为重复抽样结构,记为$(\mathscr{X},\mathscr{B},\mathscr{P})^n$或$(\mathscr{X}^n,\mathscr{B}^n,\mathscr{P}^n)$.

乘积结构在实际中相当于独立观察系统,重复抽样结构相当于对一个总体进行有限次独立抽样结果的描述.今后,"从一个总体(或分布)抽取一个样本"与"从一个统计结构抽取一个样本"这两种说法是表示同一个意思.

例 1.2 在方差相等的两个正态均值的比较的问题中,若对于第一个正态总体$N(\mu_1,\sigma^2)$获得n_1个观察值,对于第二个正态总体$N(\mu_2,\sigma^2)$获得n_2个观察值,那么研究这个问题所涉及的统计结构是一个乘积结构.

$$\begin{aligned}&(\mathbf{R}^{n_1},\mathscr{B}_{\mathbf{R}}^{n_1},\mathscr{P}_1^{n_1})\otimes(\mathbf{R}^{n_2},\mathscr{B}_{\mathbf{R}}^{n_2},\mathscr{P}_2^{n_2})\\&=(\mathbf{R}^{n_1+n_2},\mathscr{B}_{\mathbf{R}}^{n_1+n_2},\mathscr{P}_1^{n_1}\otimes\mathscr{P}_2^{n_2})\end{aligned}$$

其中

$$\mathscr{P}_1=\{N(\mu_1,\sigma^2):(\mu_1,\sigma)\in\mathbf{R}\times\mathbf{R}^+\}$$
$$\mathscr{P}_2=\{N(\mu_2,\sigma^2):(\mu_2,\sigma)\in\mathbf{R}\times\mathbf{R}^+\}$$

$(\mathbf{R}^{n_1},\mathscr{B}_{\mathbf{R}}^{n_1},\mathscr{P}_1^{n_1})$是第一个正态总体的重复抽样结构,而$(\mathbf{R}^{n_2},\mathscr{B}_{\mathbf{R}}^{n_2},\mathscr{P}_2^{n_2})$是第二个正态总体的重复抽样结构.

从定义1.2可以看出,从统计结构$(\mathscr{X},\mathscr{B},\mathscr{P})$抽取容量为$n$的样本和从重复抽样结构$(\mathscr{X}^n,\mathscr{B}^n,\mathscr{P}^n)$抽取容量为1的样本所给出的信

息是一样的.

定义 1.3 设$(\mathscr{X},\mathscr{B},\mathscr{P})^n$为重复抽样结构,对每个样本$(X_1,\cdots,X_n)\in\mathscr{X}^n$,由

$$F_n(x)=\frac{1}{n}\sum_{i=1}^{n}I_{\{X_i\leqslant x\}},\quad \forall\, x\in\mathbf{R}$$

所确定的$(\mathscr{X},\mathscr{B})$上的分布称为样本分布或经验分布.对每一个样本观察值来说,$F_n(x)$是一个分布函数,称为样本分布函数或经验分布函数.对每个固定的$x\in\mathbf{R}$,$F_n(x)$又是样本$X_1,\cdots,X_n$的一个函数,故$F_n(x)$又是一个随机变量.

由于可把诸示性函数$I_{\{X_i\leqslant x\}}$,$i=1,\cdots,n$,看作是独立同分布,仅取0或1的随机变量,故有

$$\mathrm{E}F_n(x)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}I_{\{X_i\leqslant x\}}=\frac{1}{n}\sum_{i=1}^{n}P(X_i\leqslant x)=F(x)$$

$$\begin{aligned}\mathrm{Var}[F_n(x)]&=\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}I_{\{X_i\leqslant x\}}\\&=\frac{1}{n}F(x)[1-F(x)]\leqslant\frac{1}{4n}\leqslant\frac{1}{4}\end{aligned}$$

其中$F(x)$为分布族$\mathscr{P}$中某个P的分布函数,常称$F(x)$为某总体的分布函数.由大数定律可知,对任意$\varepsilon>0$,总有

$$P(|F_n(x)-F(x)|\geqslant\varepsilon)\to 0\quad(n\to\infty)\tag{1.1}$$

这表明,只要n愈来愈大,样本的经验函数$F_n(x)$可以愈来愈接近总体分布函数$F(x)$,因此可以用$F_n(x)$的各阶矩(如样本均值,样本方差,样本相关系数,样本协方差阵等)研究统计结构$(\mathscr{X},\mathscr{B},\mathscr{P})$的某些特征.这一想法在统计中是经常被使用的.

关于经验分布函数$F_n(x)$,还有比(1.1)更强的结论,那就是如下的格里汶科定理.

定理 1.1(格里汶科) 对任意给定的自然数n,设$x_1,\cdots,x_n$是取自总体分布函数$F(x)$的一个样本观察值,$F_n(x)$为其经验分布函数,记

$$D_n=\sup_{-\infty<x<\infty}|F_n(x)-F(x)|$$

则有

$$P(\lim_{n\to\infty} D_n = 0) = 1$$

这个定理的证明在很多教科书[5]上都可找到,这里不再叙述了.容易看出,上述 D_n 可用来衡量 $F_n(x)$ 与 $F(x)$ 之间在所有的 x 值上最大差异程度.该定理表明:在 n 无限大时,对于所有的 x 值,$F_n(x)$ 与 $F(x)$ 之差的绝对值是一致地愈来愈小,这个事件发生的概率为 1.

§1.1.3 可控结构

目前对统计结构研究最多,所获结果也较多的是可控结构,为了叙述可控结构,我们从测度的绝对连续性谈起.

定义 1.4 设 μ 与 v 是可测空间 $(\mathscr{X},\mathscr{B})$ 上的两个 σ 有限测度,假如对 $\mathscr{B}$ 中使 $\mu(N)=0$ 的某个集合 N,也使 $v(N)=0$,则称 v 对 μ 是绝对连续的,或者 v 被 μ 所控制,记为 $v\ll\mu$.

根据 Radon-Nikodym 定理,在 $v\ll\mu$ 条件下,一定存在这样一个 $\mathscr{B}$ 可测函数 $p(x)$(定义在 $\mathscr{X}$ 上),使得

$$v(B) = \int_B p(x)\mu(\mathrm{d}x), \quad \forall\, B \in \mathscr{B}$$

此函数 $p(x)$ 称为 v 对 μ 的 Radom-Nikodym 导数,记为

$$p(x) = \frac{\mathrm{d}v}{\mathrm{d}\mu}, \quad \text{a.s. } [\mu]$$

并且这个函数 $p(x)$ 在 a.s. $[\mu]$ 意义下是唯一的,这里 a.s. $[\mu]$ 表示"除测度 μ 为零的集合外都成立",常称"对 μ 几乎处处成立",或称"以概率 1 成立".

定义 1.5 设 $(\mathscr{X},\mathscr{B},\mathscr{P})$ 为一统计结构,若在可测空间 $(\mathscr{X},\mathscr{B})$ 上存在这样一个 σ 有限测度 μ,使得 $\mathscr{P}$ 中每一个概率分布 P 对 μ 都是绝对连续的,即

$$P \ll \mu, \quad \forall\, P \in \mathscr{P}$$

则称该结构是可控的,相应的 Radon-Nikodym 导数 $\mathrm{d}P/\mathrm{d}\mu$ 称为概率密度函数.

在数理统计中常用来作控制的 σ 有限测度是两种:计数测度和

Lebesgue 测度.

例 1.3(计数测度) 设 $\mathscr{X}=\mathbf{R}$, $\mathscr{B}$ 是直线上一切 Borel 集组成的 σ 代数,在 $(\mathscr{X},\mathscr{B})$ 上定义如下测度

$$\mu(B)=B\text{ 中非负整数的个数},\quad \forall B\in\mathscr{B}$$

容易验证,这样的测度是 σ 有限测度,并称为计数测度.它可以控制任一个定义在非负整数集合 $\mathbf{N}$(或其子集)上的概率分布族,其 Radon-Nikodym 导数就是通常的概率分布列.如对 Poisson 分布族而言,任一个不含非负整数的 Borel 集 A 的计数测度 $\mu(A)$ 为零,而在这样的集合上 Poisson 概率 $P(A)$ 必为零,而对任一个 Borel 集 B, Poisson 概率 $P(B)$ 可表示为

$$P(B)=\int_B\frac{\lambda^x}{x!}e^{-\lambda}\mu(\mathrm{d}x)=\sum_{x\in\mathbf{B}\cap\mathbf{N}}\frac{\lambda^x}{x!}e^{-\lambda}$$

这里的 x 只能取非负整数.所以 Poisson 分布对计数测度的概率密度函数

$$P_\lambda(x)=\begin{cases}\dfrac{\lambda^x}{x!}e^{-\lambda}, & x=0,1,\cdots\\ 0, & \text{其它}\end{cases}$$

类似地可对二项分布族,负二项分布族作出解释.今后对离散分布所谈论的概率密度函数就是指该分布对计数测度的 Radon-Nikodym 导数.

例 1.4(Lebesgue 测度) 设 $\mathscr{X}=\mathbf{R}$, $\mathscr{B}$ 是直线上的一切 Borel 集组成的 σ 代数,在 $(\mathscr{X},\mathscr{B})$ 上以区间长度为基础定义的 Lebesgue 测度

$$\mu(B)=B\text{ 中不相交区间的长度之和或其极限},\forall B\in\mathscr{B}$$

容易验证,Lebesgue 测度是 σ 有限测度,它可以控制任一个定义在实数集 $\mathbf{R}$ 上的连续分布,其 Radon-Nikodym 导数就是通常的概率密度函数 $p(x)$,如对正态分布族而言,任一个 Borel 集 B 的概率总可表示为

$$P(B)=\int_B\frac{1}{\sqrt{2\pi\sigma}}e^{-(x-u)^2/2\sigma}\mathrm{d}x$$

这里的 $\mathrm{d}x=\mu(\mathrm{d}x)$ 是 Lebesgue 测度.

存在既不被计数测度控制,又不被 Lebesgue 测度控制的分布族.

下面就是这样的例子.

例 1.5(Marshall-Olkin 的二元指数分布) 有两个元件组成的串联系统可能会受到三个相互独立的 Poisson 流 Z_1, Z_2, Z_3 的冲击. 其强度分别为 $\lambda_1, \lambda_2, \lambda_3$, 即在 $(0, t)$ 内 Z_i 发生 k 次冲击的概率为

$$P(Z_i = k) = \frac{(\lambda_i t)^k}{k!} e^{-\lambda_i t}, \quad k = 0, 1, \cdots, \quad i = 1, 2, 3$$

设元件受到第一次冲击就失效. 而冲击源 Z_i 发生第一次冲击时间 T_i 服从参数为 λ_i 的指数分布. 即

$$T_i \sim F_i(t) = 1 - e^{-\lambda_i t}, \quad i = 1, 2, 3$$

又设冲击源 Z_1 仅对第一个元件冲击, Z_2 仅对第二个元件冲击, Z_3 同时对两个元件冲击. 若记两个元件的寿命分别为 X 与 Y, 则有

$$X = \min(T_1, T_3), \quad Y = \min(T_2, T_3)$$

可以算得

$$\begin{aligned}
&\overline{F}(x, y) \triangleq P(X > x, Y > y) \\
&= P(\min(T_1, T_3) > x, \min(T_2, T_3) > y) \\
&= P(T_1 > x, T_2 > y, T_3 > \max(x, y)) \\
&= \exp\{-\lambda_1 x - \lambda_2 y - \lambda_3 \max(x, y)\}, \quad x > 0, y > 0
\end{aligned}$$

这就是 Marshall-Olkin 二元指数分布, 记为 $BVE(\lambda_1, \lambda_2, \lambda_3)$. 它的联合分布函数为

$$\begin{aligned}
F(x, y) = &1 - e^{-(\lambda_1 + \lambda_3)x} - e^{-(\lambda_2 + \lambda_3)y} \\
&+ e^{-\lambda_1 x - \lambda_2 y - \lambda_3 \max(x, y)}, \quad x > 0, y > 0
\end{aligned}$$

这样一来, 我们得到一个新的参数统计结构:

$$(\mathbf{R}^+ \times \mathbf{R}^+, \mathscr{B}_{\mathbf{R}^+ \times \mathbf{R}^+}, \{BVE(\lambda_1, \lambda_2, \lambda_3), \lambda_i \in \mathbf{R}^+, i = 1, 2, 3\})$$

容易看出, 该结构不可能被计数测度控制. 下面指出, 该结构也不被 Lebesgue 测度控制, 事实上, 按最先出现的冲击源可将样本空间 $\mathscr{X} = \mathbf{R}^+ \times \mathbf{R}^+$ 划分为三部分

$$G_1 = \{Z_1 \text{ 最先出现}\} = \{(x, y) : 0 < x < y\}$$

$$G_2 = \{Z_2 \text{ 最先出现}\} = \{(x, y) : 0 < y < x\}$$

$$G_3 = \{Z_3 \text{ 最先出现}\} = \{(x, y) : 0 < x = y\}$$

通过二重积分计算得:

$$P(G_1)=\lambda_1/(\lambda_1+\lambda_2+\lambda_3)>0$$

$$P(G_2)=\lambda_2/(\lambda_1+\lambda_2+\lambda_3)>0$$

$$P(G_3)=1-P(G_1)-P(G_2)=\lambda_3/(\lambda_1+\lambda_2+\lambda_3)>0$$

区域 G_3 是半直线，它的二维 Lebesgue 测度为零，所以该结构不能被二维 Lebesgue 测度控制. 但此结构还是很有用的.

§1.2 常用分布族

在统计结构 $(\mathscr{X},\mathscr{B},\mathscr{P})$ 的三元素中，$\mathscr{X}$ 与 $\mathscr{B}$ 是不可少的，它指出应讨论哪一类事件是有意义的，但分布族 $\mathscr{P}$ 是最重要的，也是统计结构的核心. 在概率论与数理统计的一般教程中已介绍过一些分布族，它们是

二项分布族 $\{b(n,p):0<p<1\}$

Poisson 分布族 $\{P(\lambda):\lambda>0\}$

均匀分布族 $\{U(a,b):-\infty<a<b<\infty\}$

正态分布族 $\{N(\mu,\sigma^2):(\mu,\sigma)\in\mathbf{R}\times\mathbf{R}^+\}$

这些分布族及其性质都被假定是大家熟知的. 在数理统计中还将经常涉及 Gamma 分布族，Beta 分布族，Z 分布族，t 分布族，多项分布族，多元正态分布族和一些非中心分布族，本节将逐个介绍这些分布族.

§1.2.1 Gamma 分布族

定义 1.6 在 $(\mathbf{R}^+,\mathscr{B}_{\mathbf{R}^+})$ 上用密度函数

$$p(x;\alpha,\lambda)=\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x} \tag{1.2}$$

表示的概率分布称为 Gamma 分布，记为 $Ga(\alpha,\lambda)$，其中 α 与 λ 是两个正参数，α 称为形状参数，λ 称为尺度参数，Gamma 分布族常记为 $\{Ga(\alpha,\lambda):\alpha>0,\lambda>0\}$.

关于 Gamma 分布族作如下的讨论.

1. 固定尺度参数 λ 时，改变 α 将导致 Gamma 密度曲线形状的改变，图 1.1 给出 α 取不同值时几种典型的 Gamma 密度曲线，从图上可

明显看出，当 $\alpha\leqslant 1$ 时，$p(x)$是严减函数；当 $1<\alpha\leqslant 2$ 时，$p(x)$先上凸，后下凸；当 $\alpha>2$ 时，$p(x)$先下凸，后上凸，最后又下凸，此时 $p(x)$有两个拐点.

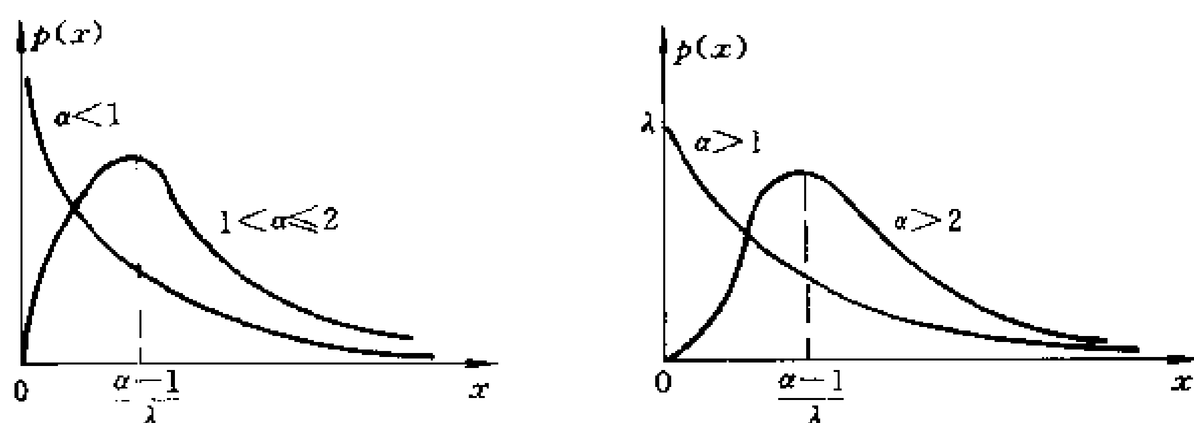


图 1.1 几种典型的 Gamma 密度曲线

2. Gamma 变量 X 的 k 阶矩为

$$EX^k = \frac{(\alpha+k-1)(\alpha+k-2)\cdots\alpha}{\lambda^k} = \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}\frac{1}{\lambda^k} \quad (1.3)$$

它的期望与方差分别为

$$EX = \frac{\alpha}{\lambda}, \mathrm{Var}(X) = \frac{\alpha}{\lambda^2} \quad (1.4)$$

3. Gamma 分布的特征函数 $f(t)$为

$$f(t) = (1 - it\lambda)^{-\alpha}$$

由此可以看出：若 $X_i \sim Ga(\alpha_i,\lambda)$，$i=1,2,\cdots,n$，且诸 X_i 间相互独立，则其和为

$$X_1 + \cdots + X_n \sim Ga(\alpha_1 + \cdots + \alpha_n, \lambda) \quad (1.5)$$

这个性质称为 Gamma 分布的可加性，这里应该强调的是“尺度参数相同”这个条件，失去这个条件，就失去了可加性.

4. Gamma 分布族有两个重要子族在数理统计中常被应用.

一是指数分布族，在 Gamma 分布中令 $\alpha=1$ 即得指数分布，记为 $Exp(\lambda)$，即 $Ga(1,\lambda)=Exp(\lambda)$. 其密度函数为

$$p(x;\lambda) = \lambda e^{-\lambda x}, \quad x > 0 \quad (1.6)$$

另一是 χ^2 分布族，在 Gamma 分布中令 $\alpha=\frac{n}{2},\lambda=\frac{1}{2}$ 即得自由度为 n 的 χ^2 分布，记为 $\chi^2(n)$，即 $Ga\left(\frac{n}{2},\frac{1}{2}\right)=\chi^2(n)$，其密度函数为

$$p(x;n)=\frac{1}{2^{n/2}\Gamma(n/2)}x^{\frac{n}{2}-1}e^{-\frac{x}{2}},\quad x>0 \tag{1.7}$$

其中自由度 n 可为任意正实数，但在实际中常用的自由度 n 为自然数，并为它们编制了 χ^2 分位数表.

§1.2.2 Beta 分布族

定义 1.7 记 $D=(0,1)$，定义在 $(D,\mathscr{B}_D)$ 上用密度函数

$$p(x;a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1} \tag{1.8}$$

表示的概率分布称为 Beta 分布，记为 $Be(a,b)$，其中 a 与 b 是两个正参数，Beta 分布族记为 $\{Be(a,b):a>0,b>0\}$.

关于 Beta 分布族作如下讨论.

1. 参数 a 与 b 的变化将导致 Beta 密度曲线形状的变化，图 1.2 给出了 a 与 b 取不同值时几种典型的 Beta 密度曲线，从图上可以看出 Beta 密度曲线 $p(x)$ 有如下几种类型.

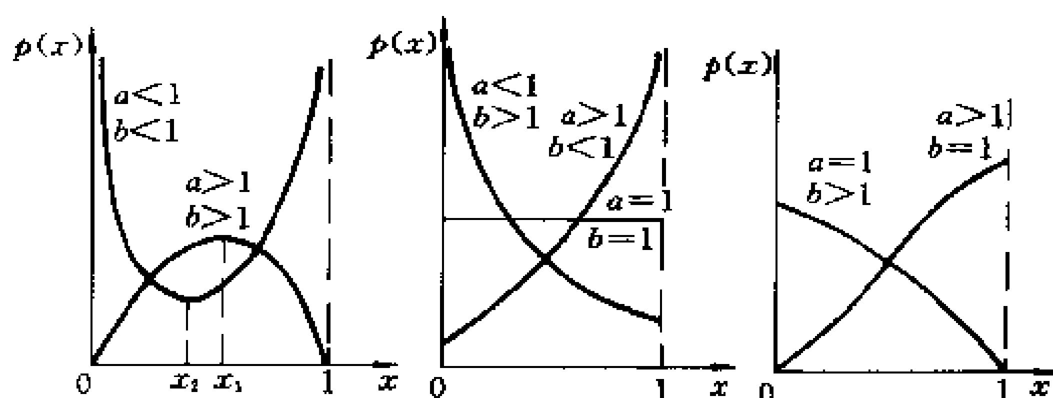


图 1.2 几种典型的 Beta 密度曲线

$a>1$ 和 $b>1$，$p(x)$ 呈单峰状，在 $x_1=(a-1)/(a+b-2)$ 处达到最大值.

$a<1$ 和 $b<1$，$p(x)$呈 U 形，在 $x_2=(1-a)/(2-a-b)$处达到最小值，当 $a=b=1/2$ 时，Beta 分布称为反正弦分布.

$a=1$ 和 $b=1$，Beta 分布就是$(0,1)$上的均匀分布，记为 $U(0,1)$，即 $Be(1,1)=U(0,1)$.

$a\leqslant 1$ 和 $b>1$，$p(x)$是严减函数.

$a>1$ 和 $b\leqslant 1$，$p(x)$是严增函数.

2. Beta 变量 X 的 k 阶矩为

$$\begin{aligned}\mathrm{E}(X^k)&=\frac{a(a+1)\cdots(a+k-1)}{(a+b)(a+b+1)\cdots(a+b+k-1)}\\&=\frac{\Gamma(a+k)\Gamma(a+b)}{\Gamma(a)\Gamma(a+b+k)}\end{aligned}\tag{1.9}$$

它的期望与方差分别为

$$\mathrm{E}X=\frac{a}{a+b},\quad \mathrm{Var}(X)=\frac{ab}{(a+b)^2(a+b+1)}\tag{1.10}$$

§1.2.3 Fisher Z 分布族

定义 1.8 定义在$(\mathbf{R}^+,\mathscr{B}_{\mathbf{R}^+})$上用密度函数

$$p(x;a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{x^{a-1}}{(1+x)^{a+b}}\tag{1.11}$$

表示的概率分布称为 Fisher Z 分布，简称 Z 分布，记为 $Z(a,b)$，其中 a 与 b 是两个正参数，Z 分布族记为$\{Z(a,b):a>0,b>0\}$.

关于 Z 分布族作如下讨论.

1. 参数 a 的变化将导致 Z 分布密度曲线形状的变化，图 1.3 给出了 a 取不同值时几种典型的 Z 分布密度曲线，从图上可以看出，当 $a\leqslant 1$时，$p(x)$是严减函数；当 $a>1$时，$p(x)$是单峰函数，且在 $x_1=(a-1)/(b+1)$处达到最大值.

2. Z 分布的 k 阶矩为

$$\mathrm{E}(X^k)=\frac{(a+k-1)(a+k-2)\cdots a}{(b-1)(b-2)\cdots(b-k)},\quad k<b\tag{1.12}$$

它的期望与方差分别为

$$\mathrm{E}(X)=\frac{a}{b-1},b>1$$

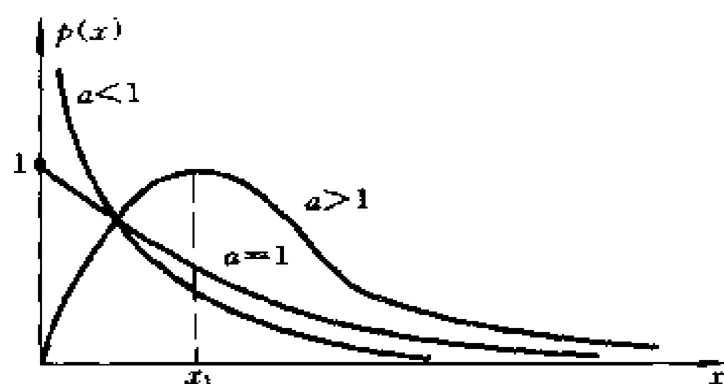


图 1.3 几种典型的 Z 分布密度曲线

$$\mathrm{Var}(X)=\frac{a(a+b-1)}{(b-1)^2(b-2)},b>2 \tag{1.13}$$

3. Z 分布与 Beta 分布之间的关系. 容易验证:

若 $X\sim Be(a,b)$,则 $Y=X/(1-X)\sim Z(a,b)$,

若 $X\sim Z(a,b)$,则 $Y=X/(1+X)\sim Be(a,b)$.

4. Z 分布与 F 分布之间的关系. 若 $X\sim Z(n_1/2,n_2/2)$,则容易导出 $Y=(n_2/n_1)X$ 的密度函数为

$$p(y;n_1,n_2)=\frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}\cdot\frac{y^{\frac{n_1}{2}-1}}{\left(1+\frac{n_1}{n_2}y\right)^{\frac{n_1+n_2}{2}}},\quad y>0 \tag{1.14}$$

这就是自由度为 n_1 和 n_2 的 F 分布,记为 $F(n_1,n_2)$,它的期望与方差分别为

$$\mathrm{E}(Y)=\frac{n_2}{n_2-2},\quad n_2>2 \tag{1.15}$$

$$\mathrm{Var}(Y)=\frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)},\quad n_2>4 \tag{1.16}$$

自由度 n_1 和 n_2 均为自然数时的 F 分布是常用统计分布之一,特编制了 F 分布分位数表. 图 1.4 给出几种 F 分布密度函数曲线.

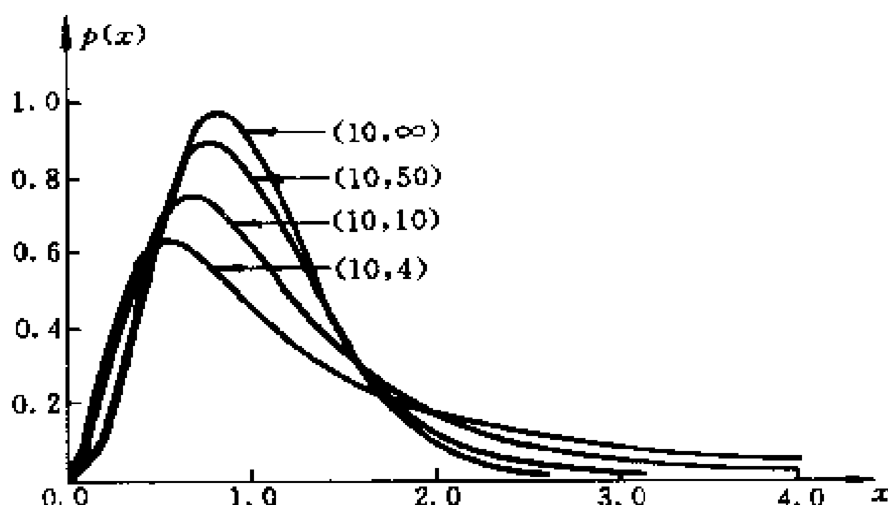


图 1.4 几种 F 分布密度曲线

§1.2.4 t 分布族

定义 1.9 定义在$(\mathbf{R},\mathscr{B}_{\mathbf{R}})$上用密度函数

$$p(x,\alpha)=\frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\sqrt{\alpha\pi}\,\Gamma\left(\frac{\alpha}{2}\right)}\left(1+\frac{x^2}{\alpha}\right)^{-\frac{(\alpha+1)}{2}} \tag{1.17}$$

表示的概率分布称为自由度为 α 的 t 分布，记为 $t(\alpha)$，参数 α 为正实数，t 分布族记为$\{t(\alpha):\alpha>0\}$.

关于 t 分布族作如下讨论：

1. t 分布的密度函数是幂函数，比较它的幂次就可看出：自由度为 α 的 t 分布只存在低于 α 阶矩，又由于 t 分布的密度函数是偶函数，故存在的奇数阶矩为零，即

$$\mathrm{E}(t^{2k+1})=0,\quad 2k+1<\alpha \tag{1.18}$$

$$\mathrm{Var}(t^{2k})=\frac{\alpha^k}{\sqrt{\pi}}\frac{\Gamma\left(\frac{\alpha}{2}-k\right)\Gamma\left(k+\frac{1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)},\quad 2k<\alpha$$

它的期望与方差分别为

$$\mathrm{E}(t)=0,\quad \mathrm{Var}(t)=\mathrm{E}(t^2)=\frac{\alpha}{\alpha-2},\quad \alpha>2 \tag{1.19}$$

2. t 分布的密度函数形状是"中间高，两边低，左右对称"(见图1.5)，很像正态分布密度函数 $\phi(x)=(2\pi)^{-\frac{1}{2}}\exp\{-x^2/2\}$，特别当 $\alpha\to\infty$ 时，有

$$\lim_{\alpha\to\infty}p(x,\alpha)=\phi(x) \tag{1.20}$$

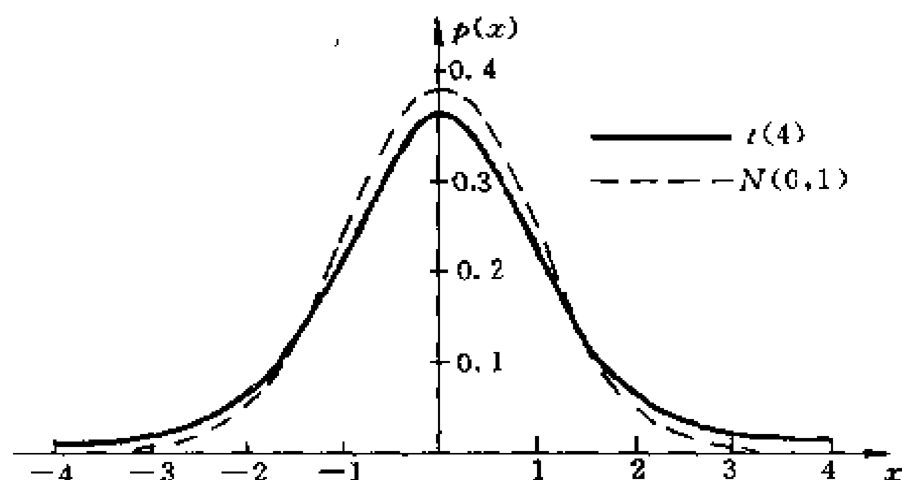


图 1.5 $t(4)$ 与 $N(0,1)$ 的密度曲线

事实上，利用 Stirling 公式可以算得

$$\lim_{\alpha\to\infty}\frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\sqrt{\alpha}\,\Gamma\left(\frac{\alpha}{2}\right)}=\frac{1}{\sqrt{2}}$$

另外，由常用的极限公式可得

$$\lim_{\alpha\to\infty}\left(1+\frac{x^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}=\lim_{\alpha\to\infty}\left[\left(1+\frac{x^2}{\alpha}\right)^{\frac{\alpha}{x^2}}\right]^{\frac{\alpha+1}{\alpha}\cdot\left(-\frac{x^2}{2}\right)}=\mathrm{e}^{-\frac{x^2}{2}}$$

综合上述即证明了(1.20)式. 实际上，当 $\alpha>30$ 时常用标准正态分布来代替 t 分布，表 1.1 列出 α 取不同值时 t 分布密度函数的最大值 $p(0;\alpha)$ 与方差的值，它们分别与标准正态密度函数的最大值 $\phi(0)=0.3989$ 与方差 1 已很接近了.

表 1.1 t 分布密度函数最大值 $p(0;\alpha)$ 与方差 $\mathrm{Var}(t)$ 随 α 的变化

α	2	4	6	10	20	30	40	∞
$p(0;\alpha)$	0.3536	0.3750	0.3827	0.3891	0.3940	0.3956	0.3965	0.3989
$\mathrm{Var}(t)$ $\frac{\alpha}{\alpha-2}$	/	2.00	1.50	1.25	1.11	1.07	1.05	1.00

3. t 分布与标准正态分布之间的微小差别是谁发现的呢？有什么价值？t 分布是英国哥塞特(Cosset,W.S. 1876—1937)在 1908 年提出的，Cosset 年轻时在英国牛津大学学习数学与化学. 1899 年在一家酿酒厂任酿酒化学技师，从事试验和数据分析工作，这项工作使他对误差有大量的感性认识，Cosset 清楚地知道，在已知总体均值 μ 和标准差 σ 时，则样本均值 $\overline{X}$ 的分布将随着样本容量 n 增大而越来越接近正态分布，但 Cosset 在试验中遇到的样本容量都不大，一般只有 5 个，他对每个样本分别计算 $\overline{X}$，S 和 t，其中

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,\quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}),\quad t = \frac{\sqrt{n}\,(\overline{x}-\mu)}{s}$$

从而获得大量 t 的观察值，发现其在$(-1,1)$，$(-2,2)$，$(-3,3)$内的频率 0.626，0.884，0.960 与 $N(0,1)$在这些区间上的概率 0.683，0.995，0.997 相差较大，于是他怀疑是否还存在一个不属于正态分布族的其它分布，他下决心研究这个问题. 在 1906 到 1907 年期间还去伦敦大学学习统计方法，与老皮尔逊(K. Pearson. 1857—1936，他的儿子 E.S. Pearson 也是著名统计学家，被人们称为小皮尔逊)共同讨论，靠着他的敏锐直觉，终于得到新的密度函数曲线，并在 1908 年以"student"笔名发表此项结果，故后人称此分布为"学生氏分布"或"t 分布". Cosset 作为统计学的新手，毅然提出一个崭新的分布是需要勇气的. 在当时，正态分布被看作是"万能分布"的时代里，代表统计学最高水平的老皮尔逊只研究大样本问题，他认为，小样本是与统计精神相违背的，是危险的倾向. 在这样的气氛下，t 分布没有被外界所理解和接受，只在 Cosset 的酿酒公司里使用. 过了一段时间后，英国另一位著名的统计

学家 Fisher 在他的农业试验中也遇到小样本问题，发现 t 分布有实用价值，直到 1923 年，Fisher 给出严格而简单的推导，1925 年又编制了 t 分布表后，Cosset 的小样本方法才被学术界承认，并获得迅速传播，发展和应用. Cosset 的 t 分布打开了人们的思路，开创了小样本方法的研究，这一段有趣的历史的更详细材料可参考文献[1].

4. 自由度为 1 的 t 分布称为柯西(Cauchy)分布，其密度函数为

$$p(x)=\frac{1}{\pi(1+x^2)},\quad x\in\mathbf{R} \tag{1.21}$$

它的更一般形式是

$$p(x;a,b)=\frac{b}{\pi[b^2+(x-a)^2]},\quad x\in\mathbf{R} \tag{1.22}$$

其中 $-\infty<a<\infty,b>0$，Cauchy 分布以它的期望与方差都不存在而著名.

§1.2.5 多项分布族

多项分布是最重要的多维离散分布. 当把一个总体按某种属性分成有限类时就会涉及这个分布. 它产生于以下的 n 次独立重复试验模型.

1. 每次试验可能有 r 种结果：$A_1,A_2,\cdots,A_r$，并且 $P(A_i)=p_i,i=1,\cdots,r,p_1+\cdots+p_r=1$.

2. 上述试验独立地重复 n 次，所得结果可用某些 A_i 组成的长为 n 的序列表示.

3. 在上述 n 次独立重复试验中，以 X_i 表示 $A_i,i=1,\cdots,r$ 出现的次数，则 $(X_1,\cdots,X_r)$ 是 r 维随机变量，在 n 次试验中 A_i 出现 $n_i,i=1,\cdots,r$ 次的概率为

$$P(X_1=n_1,\cdots,X_r=n_r)=\frac{n!}{n_1!\cdots n_r!}p_1^{n_1}\cdots p_r^{n_r} \tag{1.23}$$

其中 $n_1+\cdots+n_r=n$，诸 $n_i\geqslant 0$，这就是多项分布. 记为 $M(n,p_1,\cdots,p_r)$，当 $r=2$ 时，多项分布就退化为二项分布即 $M(n,p_1,p_2)=b(n,p_1)$.

多项分布 $M(n,p_1,\cdots,p_r)$ 的任一边际分布仍是多项分布. 譬如以 X_1 为例，它可取 $0,1,\cdots,n$ 中任一个值. 由边际分布定义可知，

$$P(X_1 = n_1) = \sum_{n_2+\cdots+n_r=n-n_1}\cdots\sum \frac{n!}{n_1!\cdots n_r!}p_1^{n_1}\cdots p_r^{n_r},\quad n_1 = 0,\cdots,n$$

若令

$$p'_2 = \frac{p_2}{1-p_1},\cdots,p'_r = \frac{p_r}{1-p_1}$$

则有 $p'_2+\cdots+p'_r=1$，若把上式改写为

$$P(X_1 = n_1) = \left(\sum_{n_2+\cdots+n_r=n-n_1}\cdots\sum \frac{(n-n_1)!}{n_2!\cdots n_r!}p'^{n_2}_2\cdots p'^{n_r}_r\right)\times$$

$$\left(\frac{n!}{n_1!(n-n_1)!}p_1^{n_1}(1-p_1)^{n-n_1}\right)$$

其中第一个括号内的和恰好是多项式$(p'_2+\cdots+p'_r)^{n-n_1}$的展开式，故其和为1，第二个括号内恰好说明 X_1 的边际分布是二项分布 $b(n,p_1)$. 这相当于把每次试验看作只有两个可能结果：A_1 和 $\overline{A}_1=A_2\cup\cdots\cup A_r$，$n$次独立重复试验中 A_1 出现 n_1 次的概率为 $P(X_1=n_1)$.

类似地，X_1 与 X_2 的边际分布是 $M=(n,p_1,p_2,p_3)$，其中 $p_3=1-p_1-p_2$，这相当于把每次试验看作只要三个可能结果：$A_1,A_2,\overline{A_1\cup A_2}=A_3\cup\cdots\cup A_r(r\geqslant4)$，$n$ 次独立重复试验中 A_1 出现 n_1 次，A_2 出现 n_2 次$(n_1+n_2\leqslant n)$的概率为

$$P(X_1 = n_1,X_2 = n_2)$$
$$= \frac{n!}{n_1!n_2!(n-n_1-n_2)!}p_1^{n_1}p_2^{n_2}(1-p_1-p_2)^{n-n_1-n_2}$$

这表明$(X_1,X_2)\sim M(n,p_1,p_2,1-p_1-p_2)$.

§1.2.6 多元正态分布族

下面我们用逐步扩充(由单元到多元，由独立到相关，由非奇异到奇异)的方式给出多元正态分布的定义.

1. 设随机变量

$$U \sim \phi(u) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-u^2/2},\quad u\in\mathbf{R}$$

则有 $\mathrm{E}(U)=0$，$\mathrm{Var}(U)=1$. 此时称 U 服从标准正态分布，记为

$N(0,1)$. 容易算得, 对任意实数 μ 和正实数 σ, 有

$$X = \mu + \sigma U \sim p(x) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbf{R}$$

并且 $\mathrm{E}(X)=\mu$, $\mathrm{Var}(X)=\sigma^2$, 这时称随机变量 X 服从期望为 μ、方差为 σ^2 的正态分布, 记为 $X \sim N(\mu,\sigma^2)$.

2. 设 $\boldsymbol{U}=(U_1,\cdots,U_n)'$, 诸 $U_i \sim N(0,1)$, 且相互独立, 则

$$\boldsymbol{U} \sim p(\boldsymbol{u}) = p(u_1,\cdots,u_n) = (2\pi)^{-\frac{n}{2}}\mathrm{e}^{-\frac{1}{2}\boldsymbol{u}'\boldsymbol{u}}$$

且 $\mathrm{E}(\boldsymbol{U})=0$, $\mathrm{Var}(\boldsymbol{U})=\boldsymbol{I}_n$, 这时称 $\boldsymbol{U}$ 服从期望为 0、协方差阵为 $\boldsymbol{I}_n$ 的 n 元标准正态分布, 记为 $\boldsymbol{U} \sim N_n(0,\boldsymbol{I}_n)$, 这里下标 n 表示分布的维数.

又设 $\boldsymbol{\mu}=(\mu_1,\cdots,\mu_n)'$ 是常数向量, $\boldsymbol{\Sigma}$ 是一个 $n\times n$ 阶正定阵, 由正定性知, 存在这样的非奇异 n 阶方阵 $\boldsymbol{A}$, 使得 $\boldsymbol{\Sigma}=\boldsymbol{A}\boldsymbol{A}'$, 于是 $\boldsymbol{\Sigma}^{-1}=\boldsymbol{A}'^{-1}\boldsymbol{A}^{-1}$, $|\boldsymbol{\Sigma}|^{-1}=|\boldsymbol{A}|^{-2}$, $|\boldsymbol{A}|^{-1}=|\boldsymbol{\Sigma}|^{-\frac{1}{2}}$. 考察随机向量

$$\boldsymbol{X} = (X_1,\cdots,X_n)' = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{U}$$

可求得 $\boldsymbol{X}$ 的联合密度函数为

$$p(\boldsymbol{x}) = (2\pi)^{-\frac{n}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}, \quad \boldsymbol{x} \in \mathbf{R}^n \qquad (1.24)$$

并且 $\mathrm{E}(\boldsymbol{X})=\boldsymbol{\mu}$, $\mathrm{Var}(\boldsymbol{X})=\boldsymbol{\Sigma}$, 这时称 $\boldsymbol{X}$ 服从期望向量为 $\boldsymbol{\mu}$, 协方差阵为 $\boldsymbol{\Sigma}$ 的非奇(非退化)多元正态分布, 记为 $\boldsymbol{X} \sim N_n(\boldsymbol{\mu},\boldsymbol{\Sigma})$.

3. 为了把非奇多元正态分布推广到更一般的情形, 即推广到协方差阵 $\boldsymbol{\Sigma}$ 是非负定情形, 此时再用密度函数工具已不行了(因为这时 $\boldsymbol{\Sigma}^{-1}$ 不存在), 因而我们要转用特征函数工具. 为此我们先计算在 $\boldsymbol{\Sigma}$ 为正定场合下多元非奇正态分布的特征函数, 对任意的向量 $\boldsymbol{t}=(t_1,\cdots,t_n)'$

$$f_{\boldsymbol{X}}(\boldsymbol{t}) = \mathrm{E}\mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}} = \mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{\mu}}\mathrm{E}\mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{A}\boldsymbol{U}}$$

令 $\boldsymbol{t}'\boldsymbol{A}=\boldsymbol{a}'=(a_1,\cdots,a_n)$, 于是

$$\begin{aligned}\mathrm{E}\mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{A}\boldsymbol{U}} = \mathrm{E}\mathrm{e}^{\mathrm{i}\boldsymbol{a}'\boldsymbol{U}} &= \prod_{i=1}^{n}\mathrm{E}\mathrm{e}^{\mathrm{i}a_iU_i} \\ &= \prod_{i=1}^{n}\mathrm{e}^{-\frac{1}{2}a_i^2} = \mathrm{e}^{-\frac{1}{2}\boldsymbol{a}'\boldsymbol{a}} \\ &= \mathrm{e}^{-\frac{1}{2}\boldsymbol{t}'\boldsymbol{A}\boldsymbol{A}'\boldsymbol{t}} = \mathrm{e}^{-\frac{1}{2}\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}}\end{aligned}$$

所以

$$f_X(t) = \exp\left\{\mathrm{i}t'\mu - \frac{1}{2}t'\Sigma t\right\} \tag{1.25}$$

我们注意到，在上述计算过程中，仅用了性质$\Sigma = AA'$，而这个性质不仅正定阵具有，即使Σ是非负定阵，也具有这个性质，差别仅在于A已不是非奇的了. 但这并不影响$f_X(t)$仍作为某一个分布的特征函数. 譬如，当Σ是非负定阵时，考察一个矩阵序列

$$\Sigma_k = \Sigma + \frac{1}{k}I_n, \quad k = 1, 2, \cdots$$

对任意的自然数k，Σ_k是正定的. 于是具有期望向量为μ，协方差阵为Σ_k的多元正态分布$N_n(\mu, \Sigma_k)$是非奇的，其特征函数

$$f_k(t) = \exp\left\{\mathrm{i}t'\mu - \frac{1}{2}t'\Sigma_k t\right\}$$

因为

$$f_X(t) = \lim_{k\to\infty} f_k(t) = \exp\left\{\mathrm{i}t'\mu - \frac{1}{2}t'\Sigma t\right\}$$

是连续函数，根据唯一性定理，$f_X(t)$一定是某n维分布的特征函数，这样的分布今后称为奇异退化n元正态分布，仍记为$N_n(\mu, \Sigma)$，一切非奇的和奇异的多元正态分布通称为多元正态分布. 这样一来，我们借助于特征函数完成了多元正态分布的定义.

定义 1.10 设$X = (X_1, \cdots, X_n)'$是一个n维随机向量，且

$\mathrm{E}(X) = \mu$（实向量），$\mathrm{Var}(X) = \Sigma$（正定阵或非负定阵）

假如它的特征函数为

$$f_X(t) = \exp\left\{\mathrm{i}t'\mu - \frac{1}{2}t'\Sigma t\right\}$$

则称X为n元正态随机向量（或称为 Gauss 随机向量），其分布称为n元正态分布，记为$X \sim N_n(\mu, \Sigma)$，而矩阵Σ的秩$\mathrm{Rank}(\Sigma) = r$称为这个分布的秩.

在这个定义中，当$\mathrm{Rank}(\Sigma) = n$时，$X$具有非奇$n$元正态分布. 其密度函数如(1.24)所示，而当$\mathrm{Rank}(\Sigma) = r < n$时，由于$\Sigma^{-1}$不存在，从而没有密度函数，那么秩为$r$的奇异$n$元正态分布到底是什么东西呢？为了回答这个问题，我们需要借助于矩阵代数的工具.

根据矩阵代数知识，对秩为 r 的非负定阵 $\boldsymbol{\Sigma}$，可以找到一个正交阵 $\boldsymbol{\Delta}$，使得

$$\boldsymbol{\Sigma} = \boldsymbol{\Delta}\begin{pmatrix} \boldsymbol{D}_\lambda & 0 \\ 0 & 0 \end{pmatrix}\boldsymbol{\Delta}'$$

其中 $\boldsymbol{D}_\lambda = \text{diag}(\lambda_1, \cdots, \lambda_r)$ 是 r 阶对角矩阵，且对角线上的元素就是 $\boldsymbol{\Sigma}$ 的 r 个非零特征根. 令 $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2)$，$\boldsymbol{\theta} = \boldsymbol{\Delta}'\boldsymbol{\mu}$，$\boldsymbol{\gamma} = \boldsymbol{\Delta}'\boldsymbol{t}$，其中 $\boldsymbol{\Delta}_1$ 是 $n\times r$ 阶阵，$\boldsymbol{\Delta}_2$ 是 $n\times(n-r)$ 阶阵，$\boldsymbol{\theta}$ 与 $\boldsymbol{\gamma}$ 是 n 维列向量，于是

$$\boldsymbol{t}'\boldsymbol{\mu} = \boldsymbol{\gamma}'\boldsymbol{\theta} = \boldsymbol{\gamma}'_1\boldsymbol{\theta}_1 + \boldsymbol{\gamma}'_2\boldsymbol{\theta}_2$$

$$\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t} = \boldsymbol{\gamma}'\begin{pmatrix} \boldsymbol{D}_\lambda & 0 \\ 0 & 0 \end{pmatrix}\boldsymbol{\gamma} = \boldsymbol{\gamma}'_1\boldsymbol{D}_\lambda\boldsymbol{\gamma}_1$$

其中 $\boldsymbol{\gamma}' = (\boldsymbol{\gamma}'_1, \boldsymbol{\gamma}'_2)$，$\boldsymbol{\theta}' = (\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2)$，$\boldsymbol{\gamma}_1$ 与 $\boldsymbol{\theta}_1$ 为 r 维列向量，因此 n 元正态向量 $\boldsymbol{X}$ 的特征函数可改写为

$$f_X(\boldsymbol{t}) = \exp\left\{\mathrm{i}\boldsymbol{\gamma}'_1\boldsymbol{\theta}_1 - \frac{1}{2}\boldsymbol{\gamma}'_1\boldsymbol{D}_\lambda\boldsymbol{\gamma}_1 + \mathrm{i}\boldsymbol{\gamma}'_2\boldsymbol{\theta}_2\right\}$$

另一方面，若令

$$\boldsymbol{y} = \boldsymbol{\Delta}'\boldsymbol{X}, \qquad \boldsymbol{y}' = (\boldsymbol{y}'_1, \boldsymbol{y}'_2), \boldsymbol{y}_1 \text{ 为 } r \text{ 维列向量}$$

由于 $\boldsymbol{\Delta}$ 是正交阵，可得

$$f_X(\boldsymbol{t}) = \mathrm{E}e^{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}} = \mathrm{E}e^{\mathrm{i}\boldsymbol{t}'\boldsymbol{\Delta}\boldsymbol{\Delta}'\boldsymbol{X}} = \mathrm{E}e^{\mathrm{i}\boldsymbol{\gamma}'\boldsymbol{y}} = \phi_y(\boldsymbol{\gamma})$$

所以有

$$f_y(\boldsymbol{t}) = \exp\left\{\mathrm{i}\boldsymbol{\gamma}'_1\boldsymbol{\theta}_1 - \frac{1}{2}\boldsymbol{\gamma}'_1\boldsymbol{D}_\lambda\boldsymbol{\gamma} + \mathrm{i}\boldsymbol{\gamma}'_2\boldsymbol{\theta}_2\right\}$$

此式对一切 $\boldsymbol{\gamma}\in\mathbf{R}^n$ 都成立. 根据特征函数的唯一性定理，上式表明随机向量 $\boldsymbol{y}_1$ 和 $\boldsymbol{y}_2$ 相互独立，并且

$$\boldsymbol{y}_1 \sim N_r(\boldsymbol{\theta}_1, \boldsymbol{D}_\lambda)$$

$$P(\boldsymbol{y}_2 = \boldsymbol{\theta}_2) = 1$$

即 $\boldsymbol{y}_2$ 的分布是退化的，而 $\boldsymbol{y}_1$ 的分布是非奇 r 元正态分布，并且这两个随机向量相互独立.

根据前面的叙述，具有非奇 r 元正态分布 $N_r(\boldsymbol{\theta}_1, \boldsymbol{D}_r)$ 的随机向量 $\boldsymbol{y}_1$ 与 $\boldsymbol{\theta}_1 + \boldsymbol{D}_\lambda^{\frac{1}{2}}\boldsymbol{U}$ 具有相同的分布，其中 $\boldsymbol{U}\sim N_r(0, \boldsymbol{I}_r)$ 于是从

$$\boldsymbol{X} = \boldsymbol{\Delta}\boldsymbol{y} = \boldsymbol{\Delta}_1\boldsymbol{y}_1 + \boldsymbol{\Delta}_2\boldsymbol{y}_2$$

可以看出，X 与 $\Delta_1\left(\theta_1+D_\lambda^{\frac{1}{2}}U\right)+\Delta_2\theta_2$ 具有相同的分布，而后者可以改写为 $\mu+BU$，其中 $B=\Delta_1D_\lambda^{\frac{1}{2}}$ 是 $n\times r$ 阶的列满秩阵，这样我们就证明了如下事实：对于任一个 n 元正态向量 $X=(X_1,\cdots,X_n)'\sim N_n(\mu,\Sigma)$，假如 Rank$(\Sigma)=r$，那么一定能找到这样一个 $n\times r$ 阶的列满秩的矩阵 B，使得 X 与 $\mu+BU$ 具有相同的分布，其中 $U\sim N_r(0,I_r)$，并且这个命题的逆命题也显然成立的．这样一来我们就得到了一个与定义 1.11 等价的定义．

定义 1.11　如果一个 n 元随机向量 $X=(X_1,\cdots,X_n)'$ 与 $\mu+BU$ 具有相同的分布，其中 μ 是一个 n 维常向量，B 是一个秩为 r 的一个 $n\times r$ 阶阵，$U\sim N_r(0,I_r)$，那么称 X 是 n 元正态向量，其分布称为 n 元正态分布，记为 $X\sim N_n(\mu,\Sigma)$，$\Sigma=BB'$，而 r 称为这个分布的秩．

多元正态分布的基本性质主要有：

1. 多元正态变量的线性变换仍然是多元正态变量．

2. 多元正态分布的低维边际分布仍然是多元或一元的正态分布．

3. 两组多元正态变量相互独立的充要条件为其协方差阵的非对角元素全为零．

4. 在多元正态变量中，一部分变量给定的条件下，另一部分变量的条件分布仍为正态分布．

这些性质都可在本章的习题中找到．只期望读者自己去证明一下．这里仅对二元正态分布作一些说明．

设 $X=(X_1,X_2)'$ 具有二元正态分布，且

$$\mathrm{E}\begin{pmatrix}X_1\\X_2\end{pmatrix}=\begin{pmatrix}\mu_1\\\mu_2\end{pmatrix}\triangleq\mu,\mathrm{Var}\begin{pmatrix}X_1\\X_2\end{pmatrix}=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\\rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\triangleq\Sigma$$

其中 $\sigma_i^2=\mathrm{Var}(X_i)$，$i=1,2$，$\rho=\mathrm{Cov}(X_1,Y_2)/(\sigma_1\sigma_2)$，$\sigma_1\sigma_2\neq0$．

显然，协方差阵 Σ 是非奇异的充要条件为 $\sigma_1\sigma_2\neq0$ 且 $|\rho|<1$．具有 $\sigma_1\sigma_2\neq0$ 且 $|\rho|<1$ 的二元正态分布的联合密度函数为

$$p(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right.\right.$$
$$\left.\left.-2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\},(x_1,x_2)\in\mathbf{R}^2$$
$$(1.26)$$

常记为$(x_1,x_2)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$.

当$\sigma_1=0$时,则有$X_1\equiv\mu_1$,$X_2\sim N(\mu_2,\sigma_2^2)$,且$X_1$与$X_2$独立.

当$\sigma_2=0$时,则有$X_2\equiv\mu_2$,$X_1\sim N(\mu_1,\sigma_1^2)$,且$X_1$与$X_2$独立.

当$\sigma_1\sigma_2\neq 0$而$|\rho|=1$时,(X_1,X_2)仍无联合密度函数可言,而其质量以概率1集中在如下一条固定的直线上.

$$\frac{X_2-\mu_2}{\sigma_2}=\rho\frac{X_1-\mu_1}{\sigma_1},\quad \rho=\pm 1$$

或者说,X_1与X_2线性相关的概率为1.

当$\sigma_1=\sigma_2=0$,这时规定$\rho=0$是合理的,而(X_1,X_2)集中在一点(μ_1,μ_2)上.

另外,当$\rho=0$时,X_1与X_2独立.反之亦然,这在一般场合下不成立.最后,在非奇异条件下容易算得如下条件分布

$$X_1|X_2=x_2\sim N\left(\mu_1+\rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2),\sigma_1^2(1-\rho^2)\right) \tag{1.27}$$

$$X_2|X_1=x_1\sim N\left(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1),\sigma_2^2(1-\rho^2)\right)$$

§1.2.7 几个非中心分布族

定义 1.12 定义在$(\mathbf{R}^+,\mathscr{B}_{\mathbf{R}^+})$上用密度函数

$$p(x;\alpha,\lambda,\gamma)=\sum_{m=0}^{\infty}\frac{e^{-\gamma}\gamma^m}{m!}p_G(x;\alpha+m,\lambda) \tag{1.28}$$

表示的概率分布称为非中心Gamma分布,记为$Ga(\alpha,\lambda,\gamma)$,其中$p_G(x;\alpha+m,\lambda)$为Gamma分布$Ga(\alpha+m,\lambda)$的密度函数.$\alpha,\lambda$是两个正实数,$\gamma$为非负实数,$\alpha$和$\lambda$仍分别称为形状参数和尺度参数,$\gamma$称为非中心参数,这里规定$0^0=1$.

显然,当$\gamma=0$时,非中心Gamma分布就退化为一般的Gamma分布,即$Ga(\alpha,\lambda,0)=Ga(\alpha,\lambda)$.

当$\alpha=n/2$和$\lambda=1/2$时,非中心Gamma分布称为非中心χ^2分布,记为$\chi^2(n,\gamma)=Ga\left(\frac{n}{2},\frac{1}{2},\gamma\right)$,显然,$\gamma=0$时,非中心$\chi^2$分布就退化为一般$\chi^2$分布,即$\chi^2(n,0)=\chi^2(n)$.

不难算出非中心 Gamma 分布 $Ga(\alpha,\lambda,\gamma)$ 的特征函数为

$$
\begin{aligned}
f(t) &= \int_0^{\infty} \mathrm{e}^{\mathrm{i}tx} p(x;\alpha,\lambda,\gamma)\mathrm{d}x \\
&= \sum_{m=0}^{\infty} \frac{\mathrm{e}^{-\gamma}\gamma^m}{m!} \frac{\lambda^{\alpha+m}}{\Gamma(\alpha+m)} \int_0^{\infty} x^{\alpha+m-1}\mathrm{e}^{-(\lambda-\mathrm{i}t)x}\mathrm{d}x \\
&= \sum_{m=0}^{\infty} \frac{\mathrm{e}^{-\gamma}\gamma^m}{m!} \left(\frac{\lambda}{\lambda-\mathrm{i}t}\right)^{\alpha+m} \\
&= \left(1-\frac{\mathrm{i}t}{\lambda}\right)^{-\alpha} \exp\frac{\mathrm{i}t\gamma}{\lambda-\mathrm{i}t}
\end{aligned}
$$

由此特征函数形式就可看出，在尺度参数 λ 相同的条件下，非中心 Gamma 分布具有可加性，即若 $X_1\sim Ga(\alpha_1,\lambda,\gamma_1)$，$X_2\sim Ga(\alpha_2,\lambda,\gamma_2)$，且 X_1 与 X_2 相互独立，则 $X_1+X_2\sim Ga(\alpha_1+\alpha_2,\lambda,\gamma_1+\gamma_2)$.

定义 1.13　定义在 $(\mathbf{R},\mathscr{B}_{\mathbf{R}})$ 上用密度函数

$$
p(x;n,\gamma) = \frac{n^{\frac{n}{2}}\mathrm{e}^{-\frac{\gamma^2}{2}}}{\sqrt{\pi}\,\Gamma\left(\frac{n}{2}\right)(n+x^2)^{\frac{n+1}{2}}} \cdot \sum_{m=0}^{\infty}\Gamma\left(\frac{n+m+1}{2}\right)\frac{\gamma^m}{m!}\left(\frac{\sqrt{2}x}{\sqrt{n+x^2}}\right)^m \tag{1.29}
$$

表示的概率分布称为自由度为 n 的非中心 t 分布，记为 $t(n,\gamma)$，其中 γ 为非负实数，称为非中心参数. 特别，当 $\gamma=0$ 时，非中心 t 分布就退化为一般的 t 分布，即 $t(n,0)=t(n)$.

§1.3　统计量及其分布

在一个实际问题归纳出的统计结构 $(\mathscr{X},\mathscr{B},\mathscr{P})$ 中，$\mathscr{X}$ 和 $\mathscr{B}$ 常可确定，而在可测空间 $(\mathscr{X},\mathscr{B})$ 上用什么分布 P 去描述尚不确定. 但我们可知道 P 属于某个分布族 $\mathscr{P}$，至于 $\mathscr{P}$ 中哪一个分布最适合还是不知道，要解决这个问题，就要从样本空间抽取样本，凭借样本中的信息对总体分布作出判断，这就是统计推断要研究的问题.

§ 1.3.1 统计量

样本中含有总体信息，但较为分散，一般不宜直接用于统计推断，常常是把样本中的信息加工处理，用样本的函数形式集中起来，这类样本函数在统计中称为统计量. 然后用统计量去作各种推断，下面先给出统计量的一般定义.

定义 1.14 设$(\mathscr{X},\mathscr{B},\mathscr{P})$是一个统计结构，$T=T(x)$是从可测空间$(\mathscr{X},\mathscr{B})$到$(\mathscr{T},\mathscr{C})$的一个可测映照，假若这个映照 T 不依赖于分布族 $\mathscr{P}$，则称 T 为此结构上的统计量，假如 $\mathscr{P}$ 为参数分布族$\{P_\theta:\theta\in\Theta\}$，则不依赖于参数 θ 的可测映照 T 称为此结构上的统计量.

在统计中样本空间常为 n 维欧氏空间，即 $\mathscr{X}=\mathbf{R}^n$，而统计量 T 的值域 $\mathscr{T}$ 常为一维欧氏空间，即 $\mathscr{T}=\mathbf{R}$，这时统计量就是不依赖于分布族（或不依赖于参数）的可测函数；假如值域 $\mathscr{T}=\mathbf{R}^k$，这时统计量就是不依赖于分布族（或不依赖于参数）的 k 个可测函数，即 $T(X)=(T_1(X),\cdots,T_k(X))$，称为向量统计量. 实际中主要使用这两种统计量.

在定义 1.14 中规定“不依赖于分布族”或“不依赖于参数”是为了得知样本 X 观察值 x 后能立即算得统计量 $T(X)$的值 $T(x)$，而不受总体分布尚未知的影响. 譬如从某样本空间 $\chi=\mathbf{R}^n$ 中抽得一个样本 $X=(X_1,\cdots,X_n)$，那么

$$\text{样本均值}\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

$$\text{样本方差}S_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

$$\text{样本偏度}\hat{\beta}_1=\frac{\sqrt{n}\sum_{i=1}^{n}(X_i-\overline{X})^3}{\left[\sum_{i=1}^{n}(X_i-\overline{X})^2\right]^{3/2}}$$

$$\text{样本峰度}\hat{\beta}_2 = \frac{n\sum_{i=1}^{n}(X_i - \overline{X})^4}{\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2} - 3$$

等都是重复抽样结构$(\mathbf{R}^n, \mathscr{B}_{\mathbf{R}}^n, \mathscr{P}^n)$上的常用统计量. 而

$$\sum_{i=1}^{n}(X_i - \mathrm{E}(X))^2, \quad \sum_{i=1}^{n}X_i^2/\mathrm{Var}(X)$$

等都不是统计量,因为总体尚未定下之前总体均值 $\mathrm{E}(X)$和总体方差 $\mathrm{Var}(X)$都是未知的,此时就是已知样本观察值也无法具体算出它们的值. 从而也无法进行统计推断.

在定义 1.14 中还强调"可测性". 这是为了保证在以后遇到与统计量 T 有关的事件时,总是有概率可言的. 如在如下的映照中

$$T: (\mathscr{X}, \mathscr{B}, \mathscr{P}) \rightarrow (\mathscr{T}, \mathscr{C})$$

谈论概率都要涉及分布族 $\mathscr{P}$,在统计结构中虽然分布尚未确定,但对 σ 代数 $\mathscr{B}$ 中任一个元素 B($\mathscr{X}$ 的某个子集,或称为事件)都可谈论概率 $P(B)$. 假如映照 T 是可测的(严格的说是 $\mathscr{B}$ 可测). 那么对 σ 代数 $\mathscr{C}$ 中任一个元素 C 的原像

$$T^{-1}(C) = \{x : T(x) \in C\} \tag{1.30}$$

一定是 σ 代数 $\mathscr{B}$ 中的元素,譬如说是 B,那么事件

$$\text{“}T \in C\text{”与“}x \in T^{-1}(C) = B\text{”}$$

是等价的. 它们发生的概率应该相等. 因此在可测性的要求下就可谈论"$T \in C$"的概率. 至于如何把概率 $P(T \in C)$计算出来那是在"抽样分布"一节中再作讨论. 这里规定可测性就是排除那些映照 T,使得原像 $T^{-1}(C)$不属于 $\mathscr{B}$. 实际中,多数的函数都是 Borel 可测函数. 可测性是一项很宽松的要求,但也是一个界限. 故在理论上从可测性来看待统计量是很重要的,必不可少的,不是任一个样本函数都可用来作统计量.

§1.3.2 抽样分布

统计量的分布称抽样分布,或称诱导分布,它在研究统计量的性质和评价一个统计推断的优良性等方面十分重要. 近代统计学的创始人

之一，英国统计学家 R. A. Fisher 曾把抽样分布，参数估计和假设检验列为统计推断的三个中心内容. 因此寻求抽样分布的理论与方法应十分重视.

设 T 是从$(\mathscr{X},\mathscr{B},\mathscr{P})$到$(\mathscr{T},\mathscr{C})$上的一个统计量. 它是样本 X 的函数. 因此对分布族 $\mathscr{P}$ 中每一个分布 P 都可确定统计量 T 的一个分布. 实际上，对任意的 $C\in\mathscr{C}$，概率

$$
\begin{aligned}
P(T(x)\in C) &= \int_{\{x: T(x)\in C\}} \mathrm{d}P \\
&= \int_{T^{-1}(C)} \mathrm{d}P = P(T^{-1}(C)) \qquad (1.31)
\end{aligned}
$$

这就是统计量 T 的分布，记为 P^T，即

$$P^T(C) = P(T^{-1}(C)),\quad \forall\, C\in\mathscr{C}$$

容易验证：这样定义的 P^T 是$(\mathscr{T},\mathscr{C})$上一个概率测度. 事实上，$P^T$ 的非负性与正则性是显然的，下面验证 P^T 的可列可加性. 设$\{C_i\}$是 $\mathscr{C}$ 中一个互不相容事件列，则总有

$$
\begin{aligned}
P^T\Big(\bigcup_{i=1}^{\infty}C_i\Big) &= P\Big(T^{-1}\Big(\bigcup_{i=1}^{\infty}C_i\Big)\Big) = P\Big(\bigcup_{i=1}^{\infty}T^{-1}(C_i)\Big) \\
&= \sum_{i=1}^{\infty}P(T^{-1}(C_i)) = \sum_{i=1}^{\infty}P^T(C_i)
\end{aligned}
$$

这个结论对分布族 $\mathscr{P}$ 中每一个分布 P 都成立. 从而我们利用统计量 T 可以诱导出另一个统计结构

$$(\mathscr{T},\mathscr{C},\mathscr{P}^T)，其中\ \mathscr{P}^T = \{P^T : P\in\mathscr{P}\}$$

这个统计结构称为 T 的诱导结构，$\mathscr{P}^T$ 称为诱导分布族.

还可证明：当原结构是可控的，那么诱导结构也是可控的. 即要证明：假如 $\mathscr{P}\ll\mu$，μ 是$(\mathscr{X},\mathscr{B})$上的 σ 有限测度，则 $\mathscr{P}^T\ll\mu^T$，其中 μ^T 是用 T 从 μ 引出的诱导测度. 首先，容易看出 μ^T 也是$(\mathscr{T},\mathscr{C})$上的 σ 有限测度，其次，对 μ^T 的零测集 N，即 $\mu^T=\mu(T^{-1}(N))=0$，由此可以推导得

$$P(T^{-1}(N)) = 0，即，P^T(N) = 0,\qquad \forall\, P\in\mathscr{P}$$

这就表明 $\mathscr{P}^T\ll\mu^T$，这也表明：原结构具有概率密度函数 $p(x)=\mathrm{d}P/\mathrm{d}\mu$，那么诱导结构也具有概率密度函数 $p(t)=\mathrm{d}P^T/\mathrm{d}\mu^T$，这也就是统计量 T

的概率密度函数，特别，原结构是参数结构，且有概率密度函数 $p(x;\theta)$，那么诱导结构也是参数结构，统计量的概率密度函数 $p(t;\theta)$ 也依赖于参数 θ.

在具体寻求统计量 T 的分布时，主要运用(1.31)，在 $\mathscr{X}$ 为欧氏空间和分布族 $\mathscr{P}$ 被 Lebesgue 测度所控的情况下，(1.31)式可改写为 Lebesgue 积分，

$$P(T(X)\in C)=\int_{\{x:T(C)\}}p(x)\mathrm{d}x \tag{1.32}$$

进一步简化(1.32)就要看统计量 T 的特征了. 一般说来，(1.32)表示的多重积分只在不多的场合下才能进一步简化，又在很少场合下才能算出明显表达式，在不能算出明显表达式的场合，可以用数值积分法获得近似分布(用分布函数表列出). 但众所周知，在三维以上的数值积分常不够准确. 下面我们列出几种特殊场合的一些结果.

1. T 是一维统计量.

设 $T=T(X_1,\cdots,X_n)$，则其分布函数为

$$F_T(t)=P(T(X_1,\cdots,X_n)\leqslant t)=\int\cdots\int_D p(x_1,\cdots,x_n)\mathrm{d}x_1\cdots\mathrm{d}x_n$$

其中积分域 $D=\{(x_1,\cdots,x_n):T(x_1,\cdots,x_n)\leqslant t\}$ 是 n 维欧氏空间的一个子集，假如 $T(X_1,\cdots,X_n)$ 是可微函数，且其梯度的模为正，即

$$\|\mathrm{grad}T(x_1,\cdots,x_n)\|^{\frac{1}{2}}=\sqrt{\sum_{i=1}^{n}\left[\frac{\partial}{\partial x_i}T(x_1,\cdots,x_n)\right]^2}>0$$

则 T 的密度函数可表示为如下 $n-1$ 维曲面积分

$$p_T(t)=\int\cdots\int_{S_{n-1}}p(x_1,\cdots,x_n)\frac{\mathrm{d}S_{n-1}}{\|\mathrm{grad}T(x_1,\cdots,x_n)\|^{\frac{1}{2}}} \tag{1.33}$$

其中积分域为方程 $T(x_1,\cdots,x_n)=t$ 所决定的 $n-1$ 维曲面 S_{n-1}.

2. $\boldsymbol{T}$ 是 k 维统计量($k<n$).

设 $\boldsymbol{T}=(T_1,\cdots,T_k)$，其中

$$T_j=T_j(X_1,\cdots,X_n),\ j=1,\cdots,k$$

是 k 个可测函数，则$(T_1,\cdots,T_k)$的联合分布函数

$$F_Y(t_1,\cdots,t_k)=\int\cdots\int_D p(x_1,\cdots,x_n)\mathrm{d}x_1,\cdots,\mathrm{d}x_n,$$

其中积分域 $D=\{(x_1,\cdots,x_n):T_j(x_1,\cdots,x_n)\leqslant t_j,j=1,\cdots,k\}$ 为 n 维欧氏空间中的一个子集，若 $T_1,\cdots,T_k$ 是可微函数，则 $(T_1,\cdots,T_k)$ 的联合密度函数可表示为如下 $n-k$ 维曲面积分

$$p_T(t_1,\cdots,t_k)=\int\cdots\int_{S_{n-k}}\frac{p(x_1,\cdots,x_n)}{\left\{\sum_{i_1<\cdots<i_k}\left[\frac{D(T_1,\cdots,T_k)}{D(x_{i_1},\cdots,x_{i_k})}\right]^2\right\}^{1/2}}\mathrm{d}S_{n-k}\quad(1.34)$$

其中积分域为由 k 个方程 $T_j(x_1,\cdots,x_n)=t_j,j=1,\cdots,k$ 所决定的 $n-k$ 维曲面 S_{n-k}，而

$$\frac{D(T_1,\cdots,T_k)}{D(x_{i_1},\cdots,x_{i_k})}=\begin{vmatrix}\frac{\partial T_1}{\partial x_{i_1}} & \cdots & \frac{\partial T_1}{\partial x_{i_k}}\\ \cdots & \cdots & \cdots\\ \frac{\partial T_k}{\partial x_{i_1}} & \cdots & \frac{\partial T_k}{\partial x_{i_k}}\end{vmatrix}$$

是函数 $T_1,\cdots,T_k$ 对变量 $x_{i_1},\cdots,x_{i_k}$ 的雅可比行列式.

3. $\boldsymbol{T}$ 是 n 维统计量.

设 $\boldsymbol{T}=(T_1,\cdots,T_n)$，其中 $T_j=T_j(X_1,\cdots,X_n)$，$j=1,\cdots,n$ 是 n 个可微函数，并存在反函数，设其反函数为

$$x_i=h_i(t_1,\cdots,t_n),i=1,\cdots,n$$

又假设这些反函数可微，则其微分元之间有如下关系

$$p(x_1,\cdots,x_n)\mathrm{d}x_1\cdots\mathrm{d}x_n=p(h_1,\cdots,h_n)|J|\mathrm{d}t_1\cdots\mathrm{d}t_n$$

其中

$$p(h_1,\cdots,h_n)=p(h_1(t_1,\cdots,t_n),\cdots,h_n(t_1,\cdots,t_n))$$

$$J=\frac{D(x_1,\cdots,x_n)}{D(t_1,\cdots,t_n)}=\left[\frac{D(t_1,\cdots,t_n)}{D(x_1,\cdots,x_n)}\right]^{-1}$$

是函数 $h_1,\cdots,h_n$ 对变量 $t_1,\cdots,t_n$ 的雅可比行列式，这时 $\boldsymbol{T}=(T_1,\cdots,T_n)$ 的联合密度函数为

$$p_T(t_1,\cdots,t_n)=p_X(h_1,\cdots,h_n)\left|\frac{D(x_1,\cdots,x_n)}{D(t_1,\cdots,t_n)}\right|\quad(1.35)$$

4. T 是 $\mathbf{R}^n$ 上的仿射变换

设 $\boldsymbol{T}=(T_1,\cdots,T_n)'$，$\boldsymbol{X}=(X_1,\cdots,X_n)'$，$\boldsymbol{A}$ 为 n 阶方阵，$\boldsymbol{C}$ 为 n 维列向量，则称 $\boldsymbol{T}=\boldsymbol{AX}+\boldsymbol{C}$ 为 $\mathbf{R}^n$ 上的仿射变换，若 $\boldsymbol{C}=0$，则 $\boldsymbol{T}$ 为 $\mathbf{R}^n$ 上的线性变换. 若 $\boldsymbol{A}$ 是非奇方阵，则其逆变换

$$\boldsymbol{X}=\boldsymbol{A}^{-1}(\boldsymbol{T}-\boldsymbol{C})$$

存在. 并且，其雅可比行列式为

$$J=\left|\det\boldsymbol{A}^{-1}\right|=|\det\boldsymbol{A}|^{-1}$$

此时，由(1.35)可得 $\boldsymbol{T}$ 的联合密度函数为

$$p_T(\boldsymbol{t})=p_X(\boldsymbol{A}^{-1}(\boldsymbol{t}-\boldsymbol{C}))/|\det\boldsymbol{A}| \tag{1.36}$$

其中 $\boldsymbol{t}=(t_1,\cdots,t_n)$. 除上述一些方法外，对特定形式的统计量还可用一些特殊技巧来获得其抽样分布，在实际应用中，常常把几种方法综合起来使用，以便更简捷地求出抽样分布，以后会经常看到这方面的应用，这里先看一个仿射变换的例子.

例 1.6 设 $X_1,\cdots,X_n$ 是来自 $N(0,1)$ 的一个样本，则可以写出 $\boldsymbol{X}=(X_1,\cdots,X_n)'$ 的联合密度函数

$$p_X(x_1,\cdots,x_n)=\frac{1}{(2\pi)^{n/2}}\exp\left\{-\frac{1}{2}\boldsymbol{x}'\boldsymbol{x}\right\}$$

其中 $\boldsymbol{x}=(x_1,\cdots,x_n)'$. 现在我们寻求 $\boldsymbol{Y}=\boldsymbol{AX}+C$ 的联合密度函数，其中 $\boldsymbol{Y}=(Y_1,\cdots,Y_n)'$，$\boldsymbol{A}$ 为 n 阶非奇矩阵，$\boldsymbol{c}=(c_1,\cdots,c_n)'$，由(1.36)可知 $\boldsymbol{Y}$ 的联合密度函数

$$\begin{aligned}p_Y(y_1,\cdots,y_n)&=\frac{1}{|\det\boldsymbol{A}|}p_X(\boldsymbol{A}^{-1}(\boldsymbol{y}-\boldsymbol{c}))\\&=\frac{1}{(2\pi)^{n/2}|\det\boldsymbol{A}|}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{c})'(\boldsymbol{AA}')^{-1}(\boldsymbol{y}-\boldsymbol{c})\right\}\end{aligned}$$

若令 $\boldsymbol{\Sigma}=\boldsymbol{AA}'$，则 $\det\boldsymbol{\Sigma}=(\det\boldsymbol{A})^2$ 或 $\det\boldsymbol{A}=(\det\boldsymbol{\Sigma})^{1/2}$. 于是

$$p_Y(y_1,\cdots,y_n)=\frac{1}{(2\pi)^{n/2}|\det\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{c})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{c})\right\}$$

这就是 n 元正态分布，其均值向量为 $\boldsymbol{C}$，方差协方差阵为 $\boldsymbol{\Sigma}$，即 $\boldsymbol{Y}\sim N_n(\boldsymbol{C},\boldsymbol{\Sigma})$. 若 $\boldsymbol{A}$ 为正交阵，则 $\boldsymbol{\Sigma}=\boldsymbol{AA}'=\boldsymbol{I}_n$，$\det\boldsymbol{\Sigma}=1$，这时，$Y_1,\cdots,Y_n$ 可看作来自 $N_n(\boldsymbol{C},\boldsymbol{I}_n)$ 的一个样本.

§1.3.3　来自正态总体的抽样分布

正态总体是实际中经常用到的一个总体.有关正态均值 μ 和正态方差 σ^2 的统计推断理论与方法贯穿于现代统计之中,所用到的统计量也十分丰富,寻求这些统计量的分布是现代统计研究的任务之一.χ^2 分布,F 分布和 t 分布是其中最重要三个抽样分布,它们的定义和基本性质已在§1.2 给出.这里应把注意力放在统计量的构造上.即从正态样本如何构造统计量才能获得这三个分布之一?其中一些明显的事实,证明就省略了.

1. 设$(X_1,\cdots,X_n)$是来自标准正态分布 $N(0,1)$的一个样本,则 $X_1^2+\cdots+X_n^2\sim\chi^2(n)$.

2. 设 $X_1,\cdots,X_n$ 是来自 $N(0,\sigma^2)$的一个样本,则

$$(X_1^2+\cdots+X_n^2)/\sigma^2\sim\chi^2(n)$$

3. 设 $X_1,\cdots,X_n$ 是来自 $N(\mu,\sigma^2)$的一个样本,则

$$(X_1^2+\cdots+X_n^2)/\sigma^2\sim\chi^2(n,\gamma)$$

其中非中心参数 $\gamma=n\mu^2/2\sigma^2$.

让我们先考察 $Y=X_1^2$ 的分布函数,当 $y\leqslant 0$ 时,$p_Y(y)=0$,当 $y>0$ 时,

$$\begin{aligned}p_Y(y)&=\left[p_Y(\sqrt{y})+p_Y(\sqrt{-y})\right]/2\sqrt{y}\\&=\frac{1}{2\sqrt{2\pi y\sigma^2}}\mathrm{e}^{-\frac{y+\mu^2}{2\sigma^2}}\left[\mathrm{e}^{\mu\sqrt{y}/2}+\mathrm{e}^{-\mu\sqrt{y}/2}\right]\\&=\frac{1}{\sqrt{2\pi y\sigma^2}}\mathrm{e}^{-\frac{y+\mu^2}{2\sigma^2}}\sum_{k=0}^{\infty}\left(\frac{\mu\sqrt{y}}{\sigma^2}\right)^{2k}/(2k)!\end{aligned}$$

由于$(2k)!=2^{2k}k!\ \Gamma\left(k+\dfrac{1}{2}\right)/\sqrt{\pi}$,故

$$p_Y(y)=\sum_{k=0}^{\infty}\frac{\left(\frac{\mu^2}{2\sigma^2}\right)^k\mathrm{e}^{-\frac{\mu^2}{2\sigma^2}}}{k!}\left(\frac{\left(\frac{1}{2\sigma^2}\right)^{k+\frac{1}{2}}}{\Gamma\left(k+\frac{1}{2}\right)}y^{k+\frac{1}{2}-1}\mathrm{e}^{-\frac{y}{2\sigma^2}}\right)$$

上式括号内不是别的，正是 Gamma 分布 $Ga\left(k+\frac{1}{2},\frac{1}{2\sigma^2}\right)$ 的密度函数，故 $Y\sim Ga\left(\frac{1}{2},\frac{1}{2\sigma^2},\frac{\mu^2}{2\sigma^2}\right)$，而 $Y/\sigma^2\sim Ga\left(\frac{1}{2},\frac{1}{2},\frac{\mu^2}{2\sigma^2}\right)=\chi^2\left(1,\frac{\mu^2}{2\sigma^2}\right)$，再由非中心 Gamma 分布的可加性得知上面第 3 个结论成立.

4. 设 $X_1,\cdots,X_n$ 是来自 $N(\mu,\sigma^2)$ 的一个样本，则统计量

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,\quad Q=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

独立，且 $\overline{X}\sim N(\mu,\sigma^2/n)$，$Q\sim\chi^2(n-1)$.

我们利用公式(1.36)来证明这个结论，为此我们先写出样本 $\boldsymbol{X}=(X_1,\cdots,X_n)'$ 的联合密度函数的矩阵形式

$$p_X(\boldsymbol{x})=\frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{x}-\mu\mathbf{1}_n)'(\boldsymbol{x}-\mu\mathbf{1}_n)\right\}$$

其中 $\mathbf{1}_n$ 为 n 个 1 组成的列向量. 接着考虑如下正交阵

$$\boldsymbol{A}=\begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 & \cdots & 0 \\ 1/\sqrt{2\times 3} & 1/\sqrt{2\times 3} & -2/\sqrt{2\times 3} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1/\sqrt{n(n-1)} & 1/\sqrt{n(n-1)} & 1/\sqrt{n(n-1)} & \cdots & (n-1)/\sqrt{n(n-1)} \\ 1/\sqrt{n} & 1/\sqrt{n} & 1/\sqrt{n} & \cdots & 1/\sqrt{n} \end{bmatrix}$$

并作如下线性变换

$$\boldsymbol{Z}=(Z_1,\cdots,Z_n)'=\boldsymbol{AX}$$

由公式(1.36)可得 $\boldsymbol{Z}$ 的联合密度函数为

$$\begin{aligned} p_Z(z)&=p_X(\boldsymbol{A}^{-1}\boldsymbol{Z})\,|\det\boldsymbol{A}|^{-1}\\ &=\frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{A}'\boldsymbol{z}-\mu\mathbf{1}_n)'(\boldsymbol{A}'\boldsymbol{z}-\mu\mathbf{1}_n)\right\}\end{aligned}$$

由正交阵性质知，$\boldsymbol{A}^{-1}=\boldsymbol{A}'$，$|\det\boldsymbol{A}|=1$，另外可算得 $\boldsymbol{Z}'\boldsymbol{A}\mathbf{1}_n=\sqrt{n}\,\mu Z_n$，于是

$$\begin{aligned}&(\boldsymbol{A}'\boldsymbol{z}-\mu\mathbf{1}_n)'(\boldsymbol{A}'\boldsymbol{z}-\mu\mathbf{1}_n)\\ &=\boldsymbol{z}'\boldsymbol{z}-\mu\boldsymbol{z}'\boldsymbol{A}\mathbf{1}_n-\mu\mathbf{1}'_n\boldsymbol{A}'\boldsymbol{z}+n\mu^2\end{aligned}$$

$$= \sum_{i=1}^{n} z_i^2 - 2\sqrt{n}\,\mu z_n + n\mu^2$$

$$= \sum_{i=1}^{n-1} z_i^2 + (z_n - \sqrt{n}\,\mu)^2$$

代回原式就可看出：$Z_1, \cdots, Z_{n-1}, Z_n - \sqrt{n}\,\mu$ 相互独立，且都服从 $N(0, \sigma^2)$，由本节的结论 2 可知

$$\frac{1}{\sigma^2}\sum_{i=1}^{n-1} Z_i^2 \sim \chi^2(n-1)$$

$$Z_n \sim N(\sqrt{n}\,\mu, \sigma^2)$$

另一方面，由正交变换可算得 $\boldsymbol{Z}'\boldsymbol{Z} = \boldsymbol{X}'\boldsymbol{X}$ 和 $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$，

$$Q = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\right)^2\right]$$

$$= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n} Z_i^2 - Z_n^2\right] = \frac{1}{\sigma^2}\sum_{i=1}^{n-1} Z_i^2 \sim \chi^2(n-1)$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{\sqrt{n}} Z_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

至于 Q 与 $\overline{X}$ 独立是由于 $Z_1, \cdots, Z_{n-1}$ 与 Z_n 相互独立所致，至此完成证明.

5. 设 $X_1, \cdots, X_{n_1}, X_{n_1+1}, \cdots, X_{n_1+n_2}$ 是来自 $N(0, \sigma^2)$ 的一个样本，则

$$F = \frac{\left(X_1^2 + \cdots + X_{n_1}^2\right)/n_1}{\left(X_{n_1-1}^2 + \cdots + X_{n_1+n_2}^2\right)/n_2} \sim F(n_1, n_2)$$

事实上，由结论 2 知

$$Q_1 = \left(X_1^2 + \cdots + X_{n_1}^2\right)/\sigma^2 \sim \chi^2(n_1) = Ga\left(\frac{n_1}{2}, \frac{1}{2}\right)$$

$$Q_2 = \left(X_{n_1+1}^2 + \cdots + X_{n_1+n_2}^2\right)/\sigma^2 \sim \chi^2(n_2) = Ga\left(\frac{n_2}{2}, \frac{1}{2}\right)$$

又从假设知 Q_1 与 Q_2 独立，再由习题 2.27 知 $Q_1/Q_2 \sim Z\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$. 最后由 §1.2.3 的第 4 个结论可知 $F = (n_2/n_1)(Q_1/Q_2) \sim F(n_1, n_2)$，这就完

成了证明.

假如把上式中 $X_1,\cdots,X_{n_1}$ 改为是来自 $N(\mu,\sigma^2)$ 的一个样本，其它条件不变，则可获得非中心 F 分布(见习题 2.17).

6. 设 $X_0,X_1,\cdots,X_n$ 是来自 $N(0,\sigma^2)$ 的一个样本，则

$$t=\frac{X_0/\sigma}{\sqrt{(X_1^2+\cdots+X_n^2)/n\sigma^2}}=\frac{X_0}{\sqrt{(X_1^2+\cdots+X_n^2)/n}}\sim t(n)$$

首先，容易看出 $t^2\sim F(1,n)$，其密度函数为

$$p_{t^2}(u)=\frac{\Gamma\left(\frac{1+n}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\cdot\frac{u^{-\frac{1}{2}}}{\left(1+\frac{u}{n}\right)^{\frac{1+n}{2}}},\quad u>0$$

其次，我们指出：当 $X_0\sim N(0,\sigma^2)$ 时，X_0 和 $-X_0$ 是同分布变量. 由此可推出 t 和 $-t$ 也是同分布变量. 于是

$$P(0<t<y)=P(0<-t<y)$$
$$=P(-y<t<0)=\frac{1}{2}P(0<t^2<y^2)$$

若记 t 的分布函数为 F_t，t^2 的分布函数为 F_{t^2}，则有

$$F_t(y)-F_t(0)=F_t(0)-F_t(-y)=\frac{1}{2}F_{t^2}(y^2)$$

对上式分别求导数，可得 t 的密度函数

$$p_t(y)=p_t(-y)=yp_{t^2}(y^2)$$

把 t^2 的密度函数代入，即得到形如(1.17)的 t 分布的密度函数.

7. 设 $X_0\sim N(\mu,\sigma^2)$，$X_1,\cdots,X_n$ 是来自 $N(0,\sigma^2)$ 的一个样本，且 X_0 与诸 X_i 相互独立，则

$$t=\frac{X_0}{\sqrt{(X_1^2+\cdots+X_n^2)/n}}\sim t(n,\gamma)$$

其中非中心参数 $\gamma=\mu/\sigma$.

这个分布亦可由定义直接导出. 具体可参阅文献[3].

§1.3.4 次序统计量及其分布

次序统计量是统计中一类常用的统计量.

定义 1.15 设 $X_1,\cdots,X_n$ 是来自某总体的一个样本. 该样本的第 i 个次序统计量,记为 $X_{(i)}$,它是如下的样本函数,每当该样本得到一组观测值 $x_1,\cdots,x_n$ 时,将它们从小到大排列为

$$x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$$

其中第 i 个值 $x_{(i)}$ 就是 $X_{(i)}$ 的观测值. 称 $(X_{(1)},\cdots,X_{(n)})$ 为该样本的次序统计量. $X_{(1)}$ 称为该样本的最小次序统计量,$X_{(n)}$ 称为该样本的最大次序统计量.

在总体有密度函数 $p(x)$ 场合,各种次序统计量的密度函数都容易用"概率元"方法导出. 大家知道,连续随机变量落在很小区间 $(x,x+\mathrm{d}x)$ 内的概率为

$$P(x < X \leqslant x + \mathrm{d}x) = p(x)\mathrm{d}x + o(\mathrm{d}x)$$

其中 $o(\mathrm{d}x)$ 是比 $\mathrm{d}x$ 高阶的无穷小量,所以 $p(x)\mathrm{d}x$ 是左端概率的主要部分,称为是 X 的概率元. 反之,若存在函数 $p(x)$ 使上式成立,则 $p(x)$ 就是 X 的密度函数. 此种寻求密度函数方法称为"概率元方法". 这个方法在多维联合密度场合也适用,下面用概率元方法来寻求各种次序统计量的密度函数.

设 $X_1,\cdots,X_n$ 是来自某总体的一个样本,该总体的分布函数为 $F(x)$,密度函数为 $p(x)$,该样本的次序统计量为

$$X_{(1)} \leqslant \cdots \leqslant X_{(n)}$$

它们的观察值依次记为 $y_1 \leqslant \cdots \leqslant y_n$.

1. $X_{(k)}$ 的密度函数 $g(y_k)$,其中 $1 \leqslant k \leqslant n$,$X_{(k)}$ 的观察值为 y_k. 以 y_k 为基础把实数轴分为三个区间:

$$(-\infty, y_k), [y_k, y_k + \mathrm{d}y_k), [y_k + \mathrm{d}y_k, \infty)$$

其中第二个区间的长度 $\mathrm{d}y_k$ 很小,使得样本观察值中只有一个落入该区间,而有两个或更多个观察值落入该区间内的概率为零或为 $o(\mathrm{d}y_k)$,这只要使 $\mathrm{d}y_k$ 充分小总可办到,这样一来,要使 $X_{(k)}$ 的观察值落入 $[y_k,y_k+\mathrm{d}y_k)$ 其内,就要样本的 n 个观察值中有 $k-1$ 个落入 $(-\infty,y_k)$

内，有 $n-k$ 个落入$[y_k+\mathrm{d}y_k,\infty)$内. 据多项分布，可算得$X_{(k)}$的概率元

$$g(y_k)\mathrm{d}y_k=\frac{n!}{(k-1)!1!(n-k)!}\times$$
$$[F(y_k)]^{k-1}p(y_k)\mathrm{d}y_k[1-F(y_k+\mathrm{d}y_k)]^{n-k}+o(\mathrm{d}y_k)$$

两边约去 $\mathrm{d}y_k$ 后，再让 $\mathrm{d}y_k\to 0$，即得 $X_{(k)}$的密度函数为

$$g(y_k)=\frac{n!}{(k-1)!(n-k)!}[F(y_k)]^{k-1}[1-F(y_k)]^{n-k}p(y_k) \tag{1.37}$$

特别，$X_{(1)}$与 $X_{(n)}$的密度函数分布为

$$g(y_1)=n[1-F(y_1)]^{n-1}p(y_1) \tag{1.38}$$

$$g(y_n)=n[F(y_n)]^{n-1}p(y_n) \tag{1.39}$$

例 1.7 设 $X_1,\cdots,X_n$ 是来自均匀分布 $U(0,1)$的一个样本，要求该样本第 k 个次序统计量 $X_{(k)}$的分布与期望.

大家知道，均匀分布 $U(0,1)$的分布函数与密度函数为

$$F(x)=\begin{cases}1, & x>1\\ x, & 0<x\leqslant 1\\ 0, & x\leqslant 0\end{cases}$$

$$p(x)=\begin{cases}1, & 0<x<1\\ 0, & \text{其它}\end{cases}$$

由公式(1.37)可得，$X_{(k)}$的密度函数为

$$g(y_k)=\frac{n!}{(k-1)!(n-k)!}x^{k-1}(1-x)^{n-k},0<y_k<1$$

其它场合 $g(y_k)=0$. 这不是别的，正是 Beta 分布 $Be(k,n-k+1)$. 由(1.11)知，$X_{(k)}$的期望为

$$\mathrm{E}(X_{(k)})=\frac{k}{n+1}$$

这个结果在使用概率纸时常用到.

2. $X_{(k)}$与 $X_{(j)}$的联合密度函数 $g(y_k,y_j)$，其中 $1\leqslant k<j\leqslant n$.

仍使用概率元方法. 以 y_k 和 y_j 分别表示 $X_{(k)}$和 $X_{(j)}$的观察值，这样就把数轴分为如下 5 个区间：

$$(-\infty,y_k),[y_k,y_k+\mathrm{d}y_k),[y_k+\mathrm{d}y_k,y_j)$$

$$[y_j + \mathrm{d}y_j), [y_j + \mathrm{d}y_j, \infty)$$

其中 $\mathrm{d}y_k$ 和 $\mathrm{d}y_j$ 都充分小，这样一来要使 $X_{(k)}$ 的观察值落入 $[y_k, y_k+\mathrm{d}y_k)$ 和 $X_{(j)}$ 的观察值落入 $[y_j, y_j+\mathrm{d}y_j)$，就必需要且仅要该样本 n 个观察值中有 $k-1$ 个落入 $(-\infty, y_k)$，有 $j-1-k$ 个落入 $[y_k+\mathrm{d}y_k, y_j)$ 和 $n-j$ 个落入 $[y_j+\mathrm{d}y_j, \infty)$，据多项分布，$X_{(k)}$ 和 $X_{(j)}$ 的概率元为

$$g(y_k, y_j)\mathrm{d}y_k\mathrm{d}y_j = \frac{n!}{(k-1)!(j-1-k)!(n-j)!} \times$$
$$[F(y_k)]^{k-1}p(y_k)\mathrm{d}y_k[F(y_j) - F(y_k + \mathrm{d}y_k)]^{j-1-k} \times$$
$$p(y_k)\mathrm{d}y_j[1 - F(y_j + \mathrm{d}y_j)]^{n-j} + o(\mathrm{d}y_k) + o(\mathrm{d}y_j)$$

两边约去 $\mathrm{d}y_k\mathrm{d}y_j$ 后，再让 $\mathrm{d}y_k$ 和 $\mathrm{d}y_j$ 都趋于零，最后得 $X_{(k)}$ 和 $X_{(j)}$ 的联合密度为

$$g(y_k, y_j) = \frac{n!}{(k-1)!(j-1-k)!(n-j)!}[F(y_k)]^{k-1} \times$$
$$[F(y_j) - F(y_k)]^{j-1-k}[1 - F(y_j)]^{n-j}p(y_k)p(y_j) \qquad (1.40)$$

这个等式在 $y_k \leqslant y_j$ 都成立，在其它场合 $g(y_k, y_j) = 0$.

用类似方法可求得任意三个或更多个次序统计量的联合密度函数，譬如：

3. 前 r 个次序统计量 $X_{(1)}, \cdots, X_{(r)}$ 的联合密度函数

$$g(y_1, \cdots, y_r) = \frac{n!}{(n-r)!}[1 - F(y_r)]^{n-r} \times$$
$$p(y_1)\cdots p(y_r), \quad y_1 \leqslant \cdots \leqslant y_n, r \leqslant n \qquad (1.41)$$

特别当 $r=n$ 时，$X_{(1)}, \cdots, X_{(n)}$ 的联合密度函数为

$$g(y_1, \cdots, y_n) = n!p(y_1)\cdots p(y_n), \quad y_1 \leqslant \cdots \leqslant y_n \qquad (1.42)$$

4. 最后，我们考虑次序统计量的矩的存在性问题.

定理 1.2　设 $X_1, \cdots, X_n$ 是来自某总体 X 的一个样本，且对某个 $a>0$ 有 $\mathrm{E}|X|^a<\infty$. 假如 n, k 和 r 满足

$$r \leqslant a \cdot \min(k, n-k+1)$$

则有 $\mathrm{E}|X_{(k)}|^r<\infty$，其中 $X_{(k)}$ 为该样本的第 k 个次序统计量.

证明：首先指出，当 $\mathrm{E}|X|^a < \infty$ 时，则对所有 $r \leqslant a$ 有 $\mathrm{E}|X|^r<+\infty$. 从而

$$\infty > \mathrm{E}|X|^r \geqslant \int_x^\infty |t|^r \mathrm{d}P(X \leqslant t) \to 0(x \to \infty)$$

另一方面，上述积分

$$\int_x^\infty |t|^r \mathrm{d}P(X \leqslant t) \geqslant |x|^r P(X > x) = |x|^r[1 - F(x)]$$

其中 $F(x)$ 为 X 的分布函数. 两边让 $x \to \infty$，可得

$$0 = \lim_{x\to\infty}\int_x^\infty |t|^r \mathrm{d}P(X \leqslant t)$$

$$\geqslant \lim_{x\to\infty} |x|^r[1 - F(x)] \geqslant 0$$

综上所述，对 $r \leqslant a$，我们有

$$|x|^r[1 - F(x)] < \infty \tag{1.43}$$

$$\lim_{x\to\infty} |x|^r[1 - F(x)] = 0 \tag{1.44}$$

其次，利用分部积分法和(1.44)

$$\infty > \mathrm{E}|X|^a \geqslant \int_0^\infty x^a \mathrm{d}P(X \leqslant x)$$

$$= \int_0^\infty x^a \mathrm{d}[1 - P(X > x)] = -\int_0^\infty x^a \mathrm{d}P(X > x)$$

$$= - x^a P(X > x)\Big|_0^\infty + a\int_0^\infty x^{a-1} P(X > x)\mathrm{d}x$$

$$= a\int_0^\infty x^{a-1}[1 - F(x)]\mathrm{d}x$$

故有

$$\int_0^\infty x^{a-1}[1 - F(x)]\mathrm{d}x < \infty \tag{1.45}$$

由第 k 个次序统计量 $X_{(k)}$ 的定义容易获得 $X_{(k)}$ 的分布函数

$$G_k(x) = P(X_{(k)} \leqslant x)$$

$$= P((X_1,\cdots,X_n) \text{ 中至少有 } k \text{ 个} \leqslant x)$$

$$= \sum_{i=k}^n \binom{n}{i}[F(x)]^i[1 - F(x)]^{n-i}$$

于是 $X_{(k)}$ 的 r 阶绝对矩可分为如下两个积分：

$$\mathrm{E}|X_{(k)}|^r = \int_{-\infty}^0 |x|^r \mathrm{d}G_k(x) + \int_0^\infty x^r \mathrm{d}G_k(x)$$

我们先考察后一个积分，用类似于获得(1.45)的方法，有

$$\int_0^{\infty} x^r \mathrm{d}G_k(x) = r\int_0^{\infty} x^{r-1}P(X_{(k)} \geqslant x)\mathrm{d}x$$

$$= r\sum_{i=0}^{k-1}\binom{n}{i}\int_0^{\infty} x^{r-1}[F(x)]^i[1-F(x)]^{n-i}\mathrm{d}x$$

$$= r\sum_{i=0}^{k-1}\binom{n}{i}\int_0^{\infty}\{x^a[1-F(x)]\}^{(r-a)/a}[1-F(x)]^{n-i-r/a}\cdot$$

$$[F(x)]^i x^{a-1}[1-F(x)]\mathrm{d}x \tag{1.46}$$

由(1.43)知$\{x^a[1-F(x)]\}^{(r-a)/a}$是有界的，从分布函数性质知$[F(x)]^i$有界，当$n-i-r/a\geqslant 0$时，$[1-F(x)]^{n-i-r/a}$亦有界，其中$i$可以为0,1,…,$k-1$. 故当$r\leqslant a(n-k+1)$时，由(1.46)和(1.45)可得

$$\int_0^{\infty} x^r \mathrm{d}G_k(x) \leqslant M\int_0^{\infty} x^{a-1}[1-F(x)]\mathrm{d}x < \infty$$

其中M是一个有限数，类似地，我们也可得到：当$r\leqslant ak$时，

$$\int_{-\infty}^0 |x|^r \mathrm{d}G_k(x) < \infty$$

综合上述两个积分存在的条件可知，当$r\leqslant a\cdot\min(k,n-k+1)$时，有$\mathrm{E}|X_{(k)}|^r<\infty$，这就完成了证明.

例 1.8　设随机变量X服从 Cauchy 分布，其密度函数为

$$p(x) = \frac{1}{\pi(1+x^2)},\quad x\in\mathbf{R}$$

大家知道$\mathrm{E}|X|^r=\infty(r\geqslant 1)$，然而对任意小的$\varepsilon>0$

$$\mathrm{E}|X|^{1-\varepsilon} = \frac{1}{\pi}\int_0^{\infty} x^{-\varepsilon}\frac{2x\mathrm{d}x}{1+x^2}$$

利用分部积分法，可得

$$\mathrm{E}|X|^{1-\varepsilon} = \frac{1}{\pi}\left[\frac{\ln(1+x^2)}{x^{\varepsilon}}\Big|_0^{\infty} + \varepsilon\int_0^{\infty} x^{-1-\varepsilon}\ln(1+x^2)\mathrm{d}x\right]$$

$$= \frac{\varepsilon}{\pi}\int_0^{\infty} x^{-1-\varepsilon}\ln(1+x^2)\mathrm{d}x$$

$$= \frac{\varepsilon}{\pi}\int_0^{\infty} x^{-1-\frac{\varepsilon}{2}}\frac{\ln(1+x^2)}{x^{\varepsilon/2}}\mathrm{d}x$$

由于$\lim\limits_{x\to\infty}\ln(1+x^2)/x^{\varepsilon/2}=0$，故存在这样的$x_0>0$和$M>0$，使得当$x>x_0$

时永远有 $\ln(1+x^2)/x^{\varepsilon/2}<M$，因此由上式可得

$$\mathrm{E}|X|^{1-\varepsilon}\leqslant\frac{\varepsilon}{\pi}\left[\int_0^{x_0}x^{-1-\varepsilon}\ln(1+x^2)\mathrm{d}x+M\int_{x_0}^{\infty}x^{-1-\varepsilon/2}\mathrm{d}x\right]<\infty$$

这表明：Cauchy 变量的 $1-\varepsilon$ 阶矩存在，其中 $\varepsilon>0$ 可以任意小.

若设 $X_1,\cdots,X_n$ 是来自上述 Cauchy 分布的一个样本. 则对样本均值 $\overline{X}$ 来说，只要 $r\geqslant1$，仍有 $\mathrm{E}|\overline{X}|^r=\infty$. 然而定理 1.2 表明其样本中位数 $X_{0.5}$ 的期望，方差或高阶矩还是有可能存在.

譬如 $n=3$ 时，$X_{0.5}=X_{(2)}$，即 $k=2$，由定理 1.2 知，对

$$r\leqslant(1-\varepsilon)\min(k,n-k+1)=(1-\varepsilon)\cdot2=2-2\varepsilon$$

有 $\mathrm{E}|X_{(2)}|^r<\infty$，由于 ε 可任意小，故对容量为 3 的 Cauchy 样本来说，其样本中位数的均值是有限的，而方差是无限的. 类似地讨论可知，容量为 5 的 Cauchy 样本中位数 $X_{(3)}$ 的均值与方差都是有限的. 当 $n=51$ 时，Cauchy 样本中位数 $X_{(26)}$ 存在低于 26 阶矩.

§1.4 统计量的近似分布

在抽样分布理论中，至今已导出的 χ^2 分布，F 分布，t 分布等为代表一批精确的抽样分布，它们的应用很广，但为数不多. 在大多数场合下的抽样分布都不易导出，或导出分布过于复杂而难于应用，这迫使人们去寻求统计量的近似分布. 这里介绍一些寻求近似分布的大样本方法. 最后介绍一种随机模拟法.

§1.4.1 从中心极限定理获得渐近分布

设 $X_1,\cdots,X_n$ 是来自某总体的一个样本，假如该总体具有均值 μ 和正的有限方差 σ^2，则由中心极限定理可知

$$P\left(\frac{X_1+\cdots+X_n-n\mu}{\sqrt{n\sigma^2}}\leqslant x\right)\longrightarrow\Phi(x),\quad n\to\infty$$

其中 $\Phi(x)$ 为标准正态分布函数. 若记 $\overline{X}$ 为样本均值，则上式可改写为

$$P\left(\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}\leqslant x\right)\longrightarrow\Phi(x),\quad n\to\infty$$

或者，在 n 较大时

$$P(\bar{X} \leqslant x) \simeq \Phi\left(\frac{\sqrt{n}(x-\mu)}{\sigma}\right)$$

这表明，即使总体不为正态分布，只要样本容量 n 较大时，$\bar{X}$ 的分布函数可用正态分布 $N(\mu,\sigma^2/n)$ 的分布函数近似，n 愈大，此种近似程度愈好，这时我们称 $N(\mu,\sigma^2/n)$ 为样本均值 $\bar{X}$ 的渐近分布，记为

$$\bar{X} \sim AN(\mu,\sigma^2/n) \tag{1.47}$$

有此渐近分布，对二阶矩存在的总体，在 n 较大时，总可获得有关样本均值 $\bar{X}$ 的一些事件概率的近似值.

中心极限定理常可如此给出样本各阶矩的渐近分布. 如在样本 k 阶矩 $\overline{X^k}=(X_1^k+\cdots+X_n^k)/n$ 中，诸 X_i^k 是独立同分布随机变量，其期望 $\mu_k=\mathrm{E}(X_1^k)$，方差 $\mathrm{Var}(X_1^k)=\mu_{2k}-\mu_k^2$，只要其方差为正，且有限，则由中心极限定理知

$$\overline{X^k} \sim AN(\mu_k,(\mu_{2k}-\mu_k^2)/n) \tag{1.48}$$

§1.4.2 随机变量序列的两种收敛性

为获得更多的渐近分布，需要讨论如下两种收敛性.

定义 1.16 设 $\{Z_n\}$ 为一随机变量序列，Z 为另一个随机变量，假如对任一个 $\varepsilon>0$，有

$$P(|Z_n-Z|\geqslant\varepsilon)\longrightarrow 0,\quad n\to\infty$$

则称随机变量序列 $\{Z_n\}$ 依概率收敛于 Z，记为 $Z_n\xrightarrow{P}Z$.

依概率收敛表示：Z_n 与 Z 相差不小于任一给定量的可能性将随着 n 增大而愈来愈小，直至趋于零. 大数定律中收敛性就是 Z 为常数 z_0 时的依概率收敛.

定义 1.17 设 $\{Z_n\}$ 为一随机变量序列，$F_n(x)$ 为 Z_n 的分布函数，Z 为另一个随机变量，其分布函数为 $F(x)$，若对 $F(x)$ 的每个连续点 x 有

$$F_n(x)\longrightarrow F(x),\quad n\to\infty$$

则称随机变量序列 $\{Z_n\}$ 依分布收敛于 Z，记为 $Z_n\xrightarrow{L}Z$ 或称分布函数

序列$\{F_n(x)\}$弱收敛于$F(x)$，记为$F_n(x)\xrightarrow{w}F(x)$.

依分布收敛是用分布函数序列的收敛性，来定义相应随机变量序列的收敛性，但可忽略$F(x)$的非连续点. 中心极限定理中的收敛性就是依分布收敛. 依分布收敛性与依概率收敛性相比是一种较弱的收敛性. 在上述记号下，有如下定理.

定理 1.3 设$Z_n\xrightarrow{P}Z$，则$Z_n\xrightarrow{L}Z$.

它的证明可在文献[4]中查得，这里略去.

下面的例子说明，定理1.3之逆不成立.

例 1.9 设Z的分布关于$x=0$对称，即其分布函数满足$F(x)=1-F(-x)$. 又设

$$Z_n=\begin{cases}Z, & n\text{ 为偶数}\\ -Z, & n\text{ 为奇数}\end{cases}\qquad n=1,2,\cdots$$

容易看出诸Z_n与Z同分布，故随机变量序列$\{Z_n\}$依分布收敛于Z，但$\{Z_n\}$不依概率收敛，因为当n为奇数时

$$|Z_n-Z|=2|Z|$$

所以

$$\begin{aligned}&P(|Z_n-Z|\geqslant\varepsilon)=P(2|Z|\geqslant\varepsilon)\\&=P\left(Z\geqslant\frac{\varepsilon}{2}\right)+P\left(Z\leqslant-\frac{\varepsilon}{2}\right)\\&=1-F(\varepsilon/2)+F(-\varepsilon/2)=2F(-\varepsilon/2)\end{aligned}$$

由于$F(0)=\frac{1}{2}$，当ε充分小时，总可假设$F(-\varepsilon/2)>1/6$，从而

$$P(|Z_n-Z|\geqslant\varepsilon)>\frac{1}{3}$$

这表明$\{Z_n\}$不能依概率收敛于Z.

但是，在Z为常数c时，两种收敛性就等价了.

定理 1.4 设c为常数，若$Z_n\xrightarrow{L}c$，则$Z_n\xrightarrow{P}c$.

证明：对任意$\varepsilon>0$，有

$$\begin{aligned}&P(|Z_n-c|\geqslant\varepsilon)\\&=1-P(Z_n<c+\varepsilon)+P(Z_n\leqslant c-\varepsilon)\end{aligned}$$

$$\leqslant 1 - F_n\left(c + \frac{\varepsilon}{2} - 0\right) + F_n(c - \varepsilon)$$

由于 $Z_n \xrightarrow{L} Z = c$，$Z$ 的分布函数 $F(x)$ 在 $c+\varepsilon/2$ 和 $c-\varepsilon$ 处是连续的，故上式右端收敛于 $1-F(c+\varepsilon/2)+F(c-\varepsilon)=0$. 这就证明了结论.

§1.4.3 几个重要的结果

下面几个定理常被用来寻求一些统计量的渐近分布.

定理 1.5(Slutsky) 设$\{Z_n\}$和$\{U_n\}$是两个随机变量序列，若

$$Z_n \xrightarrow{L} Z, U_n \xrightarrow{P} c(\text{常数})$$

则有

$$\text{(i) } Z_n + U_n \xrightarrow{L} Z + c$$

$$\text{(ii) } U_n Z_n \xrightarrow{L} cZ$$

$$\text{(iii) } Z_n/U_n \xrightarrow{L} Z/c \quad (c \neq 0)$$

证明：我们只证明(i). 因为另外两个的证明类似. 首先可考虑 $Z_n + U_n$ 的分布函数

$$\begin{aligned}
&P(Z_n + U_n \leqslant x) \\
&= P(Z_n + U_n \leqslant x, |U_n - c| \leqslant \varepsilon) + P(Z_n + U_n \leqslant x, |U_n - c| > \varepsilon) \\
&\leqslant P(Z_n \leqslant x - c + \varepsilon, |U_n - c| \leqslant \varepsilon) + P(|U_n - c| > \varepsilon) \\
&\leqslant P(Z_n \leqslant x - c + \varepsilon) + P(|U_n - c| > \varepsilon)
\end{aligned}$$

因此

$$\begin{aligned}
&\lim_{n\to\infty} \sup P(Z_n + U_n \leqslant x) \\
&\leqslant \lim_{n\to\infty} P(Z_n \leqslant x - c + \varepsilon) + \lim_{n\to\infty} P(|U_n - c| > \varepsilon) \\
&= P(Z \leqslant x - c + \varepsilon) = F(x - c + \varepsilon)
\end{aligned}$$

其中 $F(\cdot)$ 为 Z 的分布函数，类似有

$$\begin{aligned}
&P(Z_n + U_n \leqslant x) \geqslant P(Z_n + U_n \leqslant x, |U_n - c| \leqslant \varepsilon) \\
&\geqslant P(Z_n \leqslant x - c - \varepsilon, |U_n - c| \leqslant \varepsilon) \\
&= P(Z_n \leqslant x - c - \varepsilon) - P(Z_n \leqslant x - c - \varepsilon, |U_n - c| \geqslant \varepsilon)
\end{aligned}$$

$$\geqslant P(Z_n \leqslant x - c - \varepsilon) - P(|U_n - c| \geqslant \varepsilon)$$

因此

$$\liminf_{n\to\infty} P(Z_n + U_n \leqslant x)$$

$$\geqslant \lim_{n\to\infty} P(Z_n \leqslant x - c - \varepsilon) - \lim_{n\to\infty} P(|U_n - c| \geqslant \varepsilon)$$

$$\geqslant P(Z \leqslant x - c - \varepsilon) = F(x - c - \varepsilon)$$

由上述两个关系式，再考虑到 ε 的任意性，即可得

$$\lim_{n\to\infty} P(Z_n + U_n \leqslant x) = F(x - c)$$

这就意味着 $Z_n + U_n \xrightarrow{L} Z + c$. 证毕.

例 1.10 设 $\{X_n\}$ 为独立同分布的随机变量序列，其均值为 μ，方差为 σ^2，且四阶中心矩 $\mathrm{E}(X_i - \mu)^4 = v_4$ 有限，现要求样本(无偏)方差

$$S_n^2 = (n-1)^{-1} \sum\nolimits_{i=1}^n (X_i - \overline{X}_n)^2$$

的渐近分布. 其中 $\overline{X} = (X_1 + \cdots + X_n)/n$.

首先记 $\gamma^2 = \mathrm{Var}[(X_i - \mu)^2] = v_4 - \sigma^4$，并记

$$\frac{S_n^2 - \sigma^2}{\gamma\sqrt{n}} = \frac{n}{n-1} \frac{1}{\gamma\sqrt{n}} (T_n - (n-1)\sigma^2)$$

其中

$$T_n = \sum_{i=1}^n (X_i - \overline{X}_n)^2 = \sum_{i=1}^n \left\{ (X_i - \mu)^2 - \frac{\sigma^2}{n} \right\} - n(\overline{X}_n - \mu)^2 + \sigma^2$$

若令

$$G_n = \sum_{i=1}^n g(X_i), \quad g(X_i) = (X_i - \mu)^2 - \sigma^2/n$$

$$R_n = -n(\overline{X}_n - \mu)^2 + \sigma^2$$

则有 $T_n = G_n + R_n$. 对 R_n 我们记 $U_n = |R_n/\sqrt{n}|$，其分布函数记为 $F_n(u)$，于是

$$P(|R_n/\sqrt{n}| \geqslant \varepsilon) = P(U_n \geqslant \varepsilon)$$

$$= \int_\varepsilon^\infty \mathrm{d}F_n(u) \leqslant \int_\varepsilon^\infty \frac{u}{\varepsilon} \mathrm{d}F_n(u)$$

$$\leqslant \frac{1}{\varepsilon} \int_0^\infty u \mathrm{d}F_n(u) = \frac{\mathrm{E}(U_n)}{\varepsilon}$$

由于

$$E(U_n) = E\left|R_n/\sqrt{n}\right| \leqslant \sqrt{n}E(\overline{X}_n - \mu)^2 + \sigma^2/\sqrt{n}$$
$$= 2\sigma^2/\sqrt{n} \longrightarrow 0 \quad (n \to \infty)$$

所以可得 $R_n/\sqrt{n} \xrightarrow{P} 0$，另外，对 G_n 我们知道 $g(X_i)$ 是独立同分布随机变量，且 $E[g(X_i)] = (n-1)\sigma^2/n$，$\mathrm{Var}[g(X_i)] = \gamma^2 < +\infty$，故由中心极限定理知

$$\frac{G_n - (n-1)\sigma^2}{\gamma\sqrt{n}} \xrightarrow{L} N(0,1)$$

这里和今后还要用 $N(0,1)$ 表示标准正态变量. 最后，由定理 1.5 可得

$$\frac{T_n - (n-1)\sigma^2}{\gamma\sqrt{n}} \xrightarrow{L} N(0,1)$$

从而

$$\frac{S_n^2 - \sigma^2}{\gamma/\sqrt{n}} \xrightarrow{L} N(0,1)$$

这说明 S_n^2 的渐近分布为 $N(\sigma^2, \gamma^2/n)$，其中 $\gamma^2 = v_4 - \sigma^4$.

定理 1.6　设 $\{Z_n\}$ 为一随机变量序列，且 $Z_n \xrightarrow{P} c$（常数），又函数 $g(\cdot)$ 是在点 c 处连续，则 $g(Z_n) \xrightarrow{P} g(c)$.

证明：由连续性可知，对任意 ε，存在 $\delta > 0$，使得当 $|z-c| < \delta$ 时，总有 $|g(z) - g(c)| < \varepsilon$，从而

$$\begin{aligned} & P(|g(Z_n) - g(c)| < \varepsilon) \\ \geqslant\ & P(|Z_n - c| < \delta) \\ =\ & 1 - P(|Z_n - c| \geqslant \delta) \longrightarrow 1 \quad (n \to \infty) \end{aligned}$$

这就说明：$g(Z_n) \xrightarrow{P} g(c)$. 证毕.

例 1.11　设 $\{X_n\}$ 是独立同分布随机变量序列，其均值为 μ，方差为 $\sigma^2 < \infty$. 模仿正态总体下的 t 统计量，构造

$$t_n = \sqrt{n}(\overline{X}_n - \mu)/S_n$$

常称 t_n 为 t 化统计量，现要求 t_n 的渐近分布.

首先由辛钦大数定律可知

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 \xrightarrow{P} \sigma^2, \quad \overline{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i \xrightarrow{P} \mu$$

因此由 Slutsky 定理可得

$$S_n^2=\frac{n}{n-1}\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-(\overline{X}_n-\mu)^2\right\} \xrightarrow{P} \sigma^2$$

再由定理 1.6 可知 $S_n \xrightarrow{P} \sigma$，另一方面，由中心极限定理可知

$$Z_n=\sqrt{n}(\overline{X}_n-\mu)/\sigma \xrightarrow{L} N(0,1)$$

最后由 Slutsky 定理即可知 $t_n \xrightarrow{L} N(0,1)$.

定理 1.7 设 $\{a_n\}$ 为一趋于 ∞ 的数列，b 为常数，并且对随机变量序列 $\{Z_n\}$ 有

$$a_n(Z_n-b) \xrightarrow{L} Z$$

又设 $g(\cdot)$ 为可微函数，且 g' 在点 b 处连续，则有

$$a_n[g(Z_n)-g(b)] \xrightarrow{L} g'(b)Z$$

证明：首先由定理的假设可得

$$Z_n-b=\frac{1}{a_n}[a_n(Z_n-b)] \xrightarrow{L} 0\cdot Z=0$$

再由定理 1.4 可知 $|Z_n-b| \xrightarrow{P} 0$.

另一方面，由中值定理知

$$a_n[g(Z_n)-g(b)]=a_n g'(Z_n^*)(Z_n-b)$$

其中 Z_n^* 满足 $|Z_n^*-b|\leqslant|Z_n-b|$. 在这个不等式中，由于右端依概率收敛于 0，故左端也必依概率收敛于 0，从而

$$Z_n^* \xrightarrow{P} b$$

再由 g' 在点 b 处连续和定理 1.6 可得

$$g'(Z_n^*) \xrightarrow{P} g'(b)$$

综合上述，再一次应用定理 1.5 即可得

$$g'(Z_n^*)[a_n(Z_n-b)] \xrightarrow{L} g'(b)Z$$

这就完成了定理的证明.

例 1.12　设$\{X_n\}$是独立同分布随机变量序列，其均值为μ、方差为σ^2，且$\gamma_4=E(x-\mu)^4<\infty$，在例 1.11 中已有

$$\sqrt{n}\,(S_n^2-\sigma^2)/\gamma \xrightarrow{L} N(0,1)$$

如今我们求$\sqrt{n}\,(S_n-\sigma)/\gamma$的渐近分布，其中$\gamma^2=\gamma_4-\sigma^4$.

虽然$S_n=\left[\sum_{i=1}^n(X_i-\overline{X})^2/(n-1)\right]^{1/2}$不能表示为独立随机变量之和，但是$S_n=\sqrt{S_n^2}$，若取$g(z)=\sqrt{z}$，则$g'(z)=(2\sqrt{z})^{-1}$再取$a_n=\sqrt{n}/\gamma$. 应用定理 1.7 即得

$$\sqrt{n}\,(S_n-\sigma)/\gamma \xrightarrow{L} (2\sigma)^{-1}\cdot Z$$

其中Z为标准正态变量，或者

$$S_n\sim AN(\sigma,\gamma^2/4n\sigma^2)$$

例 1.13　设$X_1,\cdots,X_n$是来自二点分布$b(1,\theta)$的一个样本. 其中$0<\theta<1$，现要求样本均值$\overline{X}$的函数$g(\overline{X})=\overline{X}(1-\overline{X})$的渐近分布.

由于二点分布$b(1,\theta)$的方差$\sigma^2=\theta(1-\theta)\leqslant 0.25$，故由中心极限定理知

$$\sqrt{n}\,(\overline{X}-\theta)/\sqrt{\theta(1-\theta)} \xrightarrow{L} N(0,1)$$

另外取$g(z)=z(1-z)$，于是$g'(z)=1-2z$在$(0,1)$上连续，若取$a_n=\sqrt{n}/\sqrt{\theta(1-\theta)}$，则定理 1.7 可得

$$a_n[\overline{X}(1-\overline{X})-\theta(1-\theta)] \xrightarrow{L} (1-2\theta)Z$$

其中Z为标准正态变量，只要$\theta\neq 1/2$，在n较大时有

$$\overline{X}(1-\overline{X})\sim AN\left(\theta(1-\theta),\frac{\theta(1-\theta)(1-2\theta)^2}{n}\right)$$

§1.4.4　样本的 p 分位数及其渐近分布

1. 总体的p分位数

设总体X的分布函数为$F(x)=P(X\leqslant x)$，对给定的$p\in(0,1)$，方程$F(x)=p$的解$x=\xi_p$称为该总体（或该分布）的p分位数，当$p=0.5$时，$\xi_{0.5}$称为该总体的中位数，在这个定义下可能会出现如下三种

情况[7]：

(i) 唯一的 p 分位数，如 $F(x)$ 在 ξ_p 处连续单调，见图 1.6(a).

(ii) 没有 p 分位数，如 $F(x)$ 在 ξ_p 处恰有跳跃，见图 1.6(b).

(iii) 不止一个 p 分位数，如在区间 (a,b) 上任一点 x 都有 $F(x)=p$，见图 1.6(c).

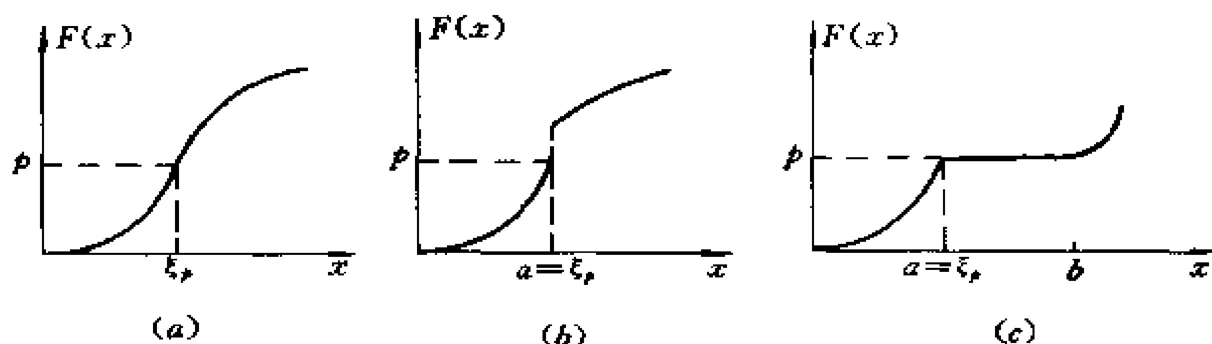


图 1.6　总体的 p 分位数

这种定义简单明确，常在实际中被采用. 但由于其存在性和唯一性问题使理论研究发生技术上的困难. 为了回避这些技术困难，统计学家又把总体的 p 分位数定义为

$$\xi_p = \inf\{x : F(x) \geqslant p\}, p \in (0,1) \tag{1.49}$$

当 $p=0.5$ 时，$\xi_{0.5}$ 称为该总体的中位数. 在这个定义下，上述三种情况都有唯一的 p 分位数，特别，当 $F(x)$ 是严增函数时，上述两种定义是等价的，常见的连续分布都属于这种情况.

2. 样本的 p 分位数.

设 $X_1,\cdots,X_n$ 是来自某总体的一个样本，$X_{(1)}\leqslant\cdots\leqslant X_{(n)}$ 是其次序统计量. 该样本的中位数定义为

$$m_{0.5} = \begin{cases} X_{\left(\frac{n+1}{2}\right)}, & \text{当 } n \text{ 为奇数} \\ \left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right)\Big/2, & \text{当 } n \text{ 为偶数} \end{cases} \tag{1.50}$$

对一般的样本的 p 分位数 m_p 虽然仍是用次序统计量给定，但给定的方式也有多种. 譬如

$$\text{(i) } m_p^{(1)} = X_{(k)}, \text{其中 } k = [np] + 1 \tag{1.51}$$

$$\text{(ii)}\ m_p^{(2)} = X_{(k)}, \text{其中}\ k = [(n+1)p] \tag{1.52}$$

$$\text{(iii)}\ m_p^{(3)} = \inf\{x : F_n(x) \geqslant p\} \tag{1.53}$$

其中$[a]$为不超过 a 的最大整数. $F_n(x)=\frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \leqslant x\}}$为样本的经验分布函数. 这几种样本分位数的定义虽有所不同，但都可用某个 $X_{(k)}$表示，其中 k 与 np 的距离一般不会超过 1. 譬如，在 $p=0.25$，$n=70$ 时可导出 $m_p^{(1)}=X_{(18)}$，$m_p^{(2)}=X_{(17)}$，$m_p^{(3)}=X_{(18)}$，这些次序统计量的下标 k 与 $np=17.50$ 相差都不超过 1. 当样本容量 n 无限增大时，k 与 np 之间的差与 n 相比，或与 $\sqrt{n}$ 相比都微不足道.

3. 样本分位数的渐近分布.

定理 1.8 设 $X_1,\cdots,X_n$ 是来自具有密度函数 $p(x)$的总体的一个样本，对给定的 $p\in(0,1)$，$p(x)$在总体的 p 分位数 ξ_p 处连续，且 $p(\xi_p)>0$，又定义 k，使得 $k=np+o(\sqrt{n})$，则对样本的第 k 个次序统计量 $X_{(k)}$有

$$\frac{\sqrt{n}(X_{(k)}-\xi_p)}{\sqrt{p(1-p)}/p(\xi_p)} \xrightarrow{L} N(0,1)$$

证明：令 $Z_n=\sqrt{n}(X_{(k)}-\xi_p)$，则有

$$P(Z_n \leqslant x) = P(\sqrt{n}(X_{(k)}-\xi_p) \leqslant x)$$

$$= P\left(X_{(k)} \leqslant \xi_p + \frac{x}{\sqrt{n}}\right)$$

$$= P\left(\sum_{i=1}^{n} I_{\{X_i \leqslant \xi_p + \frac{x}{\sqrt{n}}\}} \geqslant k\right)$$

最后一个等号成立是由于事件“$X_{(k)}\leqslant\xi_p+\frac{x}{\sqrt{n}}$”等价于“样本中至少有 k 个样品观察值不超过 $\xi_p+\frac{x}{\sqrt{n}}$”，若令

$$Y_{ni} = I_{\{X_i \leqslant \xi_p + x/\sqrt{n}\}} - F\left(\xi_p + \frac{x}{\sqrt{n}}\right), i=1,\cdots,n$$

$$t_n = \{k - nF(\xi_p + x/\sqrt{n})\}/\sqrt{n}$$

则事件"$\sum_{i=1}^{n} I_{\{X_i \leqslant \xi_p + \frac{x}{\sqrt{n}}\}} \geqslant k$"等价于事件"$\sum_{i=1}^{n} Y_{ni} \geqslant \sqrt{n}\, t_n$"，其中诸 Y_{ni} 是相互独立同分布随机变量，其分布为

$$P\left[Y_{ni} = 1 - F(\xi_p + x/\sqrt{n})\right] = F(\xi_p + x/\sqrt{n})$$

$$P\left[Y_{ni} = -F(\xi_p + x/\sqrt{n})\right] = 1 - F(\xi_p + x/\sqrt{n})$$

并且

$$\begin{aligned} F(\xi_p + x/\sqrt{n}) &= F(\xi_p) + p(\xi_p)x/\sqrt{n} + o(1/\sqrt{n}) \\ &= p + p(\xi_p)x/\sqrt{n} + o(1/\sqrt{n}) \end{aligned}$$

容易算得：$\mathrm{E}(Y_{ni}) = 0$，

$$\begin{aligned} \mathrm{Var}(Y_{ni}) = \mathrm{E}(Y_{ni}^2) &= F(\xi_p + x/\sqrt{n})\left[1 - F(\xi_p + x/\sqrt{n})\right] \\ &= p(1-p) + o(1/\sqrt{n}) \end{aligned}$$

$$\begin{aligned} t_n &= \frac{1}{\sqrt{n}}\left[k - np - np(\xi_p)x/\sqrt{n} + o(\sqrt{n})\right] \\ &= -xp(\xi_p) + o(1) \end{aligned}$$

利用中心极限定理可得

$$P(Z_n \leqslant x) = P\left(\frac{1}{\sqrt{np(1-p)}}\sum_{i=1}^{n} Y_{ni} \geqslant \frac{t_n}{\sqrt{p(1-p)}}\right)$$

$$\longrightarrow 1 - \Phi\left(\frac{-xp(\xi_p)}{\sqrt{p(1-p)}}\right) = \Phi\left(\frac{xp(\xi_p)}{\sqrt{p(1-p)}}\right)$$

这就意味着

$$P\left\{\frac{\sqrt{n}\,(X_{(k)} - \xi_p)}{\sqrt{p(1-p)}/p(\xi_p)} \leqslant x\right\} \longrightarrow \Phi(x)\,(n \to \infty)$$

这就完成了证明.

从样本的 p 分位数的几种定义看，所用的 $X_{(k)}$ 都能使 $(k-np)/\sqrt{n} \to 0(n \to \infty)$，即 $k = np + o(\sqrt{n})$，所以由定理 1.7 可得样本的 p 分位数的渐近分布

$$m_p \sim AN\left(\xi_p, \frac{p(1-p)}{n[p(\xi_p)]^2}\right) \tag{1.54}$$

特别，当 $p = 0.5$ 时，样本中位数的渐近分布为

$$m_{0.5} \sim AN\left(\xi_{0.5}, \frac{1}{4n[p(\xi_{0.5})]^2}\right) \tag{1.55}$$

§1.5 充分统计量

§1.5.1 统计量的压缩数据功能

定义在可测空间$(\mathscr{X},\mathscr{B})$上的统计量$T=T(x)$，实际上是对样本$X=(X_1,\cdots,X_n)$进行某种加工的结果. 譬如，为了估计总体的均值，人们把样本X加工为样本均值$\overline{X}$，为了估计总体方差σ^2，人们把样本X加工为样本方差S_n^2，然后用$\overline{X}$和S_n^2分别去估计均值μ和方差σ^2，这种加工本质上是统计量的压缩数据功能的体现. 如何理解这个含义呢？

直观上看，样本的不同的观察值，统计量T有相同的值，譬如，改变样本观察的排列顺序，不会改变$\overline{X}$与S_n^2的值. 这就是统计量的"压缩数据"功能. 从理论上看，若T是在$(\mathscr{T},\mathscr{C})$上取值的可测映照. 那么对$\sigma$代数$\mathscr{C}$中任一元素$C$在$\mathscr{B}$中有一个原像

$$T^{-1}(C) = \{x : T(x) \in C\} \in \mathscr{B}$$

把所有原像的全体记为

$$T^{-1}(\mathscr{C}) = \{T^{-1}(C) : C \in \mathscr{C}\} \subset \mathscr{B}$$

容易验证：$T^{-1}(\mathscr{C})$是σ代数，并且是$\mathscr{B}$的子σ代数. 这表明，有了统计量T之后，原先样本所涉及的可测空间$(\mathscr{X},\mathscr{B})$换为另一个新的可测空间$(\mathscr{X},T^{-1}(\mathscr{C}))$，差别在于$\sigma$代数缩小了. 涉及到的事件减少了. 只有在$T$是一对一映照时$\sigma$代数才不能被压缩. 而统计中所用的统计量$T=T(x)$常常是多对一映照. 所以$\sigma$代数缩小了是几乎肯定的. 这就是"压缩数据"的含义.

例 1.14 设$X=(X_1,X_2,X_3)$是从二点分布$b(1,\theta)$中抽取的一个容量为 3 的样本. 由于每个X_i仅能取 0 或 1，故其样本空间$\mathscr{X}$只含有 8 个样本点，$\mathscr{X}=\{x^{(1)},\cdots,x^{(8)}\}$，其中

$$x^{(1)} = (0,0,0),\quad x^{(2)} = (0,0,1),\quad x^{(3)} = (0,1,0)$$
$$x^{(4)} = (1,0,0)\quad x^{(5)} = (0,1,1),\quad x^{(6)} = (1,0,1)$$

$$x^{(7)} = (1,1,0), \quad x^{(8)} = (1,1,1)$$

在其上的 σ 代数 $\mathscr{B}$ 是其子集全体，共有 $2^8=256$ 个事件组成.

若在$(\mathscr{X},\mathscr{B})$上定义统计量 $T=X_1+X_2+X_3$，它的值域为 $\mathscr{T}$，仅含 4 个点，这是由于不同的样本点对应相同的 T 值之故. 这时 $\mathscr{T}$ 上的 σ 代数 $\mathscr{C}$ 也只含有 $2^4=16$ 个事件. 这 16 个事件的原像组成 $\mathscr{X}$ 上的一个 σ 代数 $T^{-1}(\mathscr{C})$. 若记 $C_i=\{T=i\}$，$i=0,1,2,3$，则诸 C_i 的原像为

$$B_0 = T^{-1}(C_0) = \{x^{(1)}\}$$

$$B_1 = T^{-1}(C_1) = \{x^{(2)},x^{(3)},x^{(4)}\}$$

$$B_2 = T^{-1}(C_2) = \{x^{(5)},x^{(6)},x^{(7)}\}$$

$$B_3 = T^{-1}(C_3) = \{x^{(8)}\}$$

这 4 个原像实际上是 $\mathscr{X}$ 的一个分割，这个分割产生的 σ 代数就是 $T^{-1}(\mathscr{C})$，显然 $T^{-1}(\mathscr{C})\subset\mathscr{B}$，即 σ 代数中的元素由 256 个减少至 16 个.

一般说来，任一个统计量都有压缩数据的功能，只是程度不同罢了. 譬如，如下两个统计量

$$T_1 = X_1X_2X_3, \quad T_2 = a(\text{常数})$$

容易看出，它们产生的子 σ 代数分别为

$$T_1^{-1}(\mathscr{C}_1) = \{\varnothing,\{x^{(1)},\cdots,x^{(7)}\},\{x^{(8)}\},\mathscr{X}\}$$

$$T_2^{-1}(\mathscr{C}_2) = \{\varnothing,\mathscr{X}\}$$

其中 $\varnothing$ 为空集（不可能事件），$\mathscr{C}_i$ 为 T_i 值域上的 σ 代数. 这两个统计量的压缩数据功能并不亚于统计量 T. 但在实际中，这些统计量不被采用. 原因是压缩过分了，以至于抹杀了样本之间的重要差别，造成信息损失. 因为样本 $X=(x_1,x_2,x_3)$提供了两种信息：

1. 三次试验中成功出现多少次.

2. 成功出现在那些试验点上.

对估计成功概率 θ 来说，有用的量是第一种信息，第二种信息对估计 θ 并不重要. 统计量 T_1 和 T_2 没有把第一种信息完全反映出来，而统计量 $T=X_1+X_2+X_3$ 把第一种信息充分反映出来了. 所以说统计量 T 既压缩数据，又不损失有关参数 θ 的信息，而统计量 T_1 和 T_2 虽然压缩数据，但损失了有关参数 θ 的信息，在统计中把不损失信息的统计量称为

充分统计量.

§1.5.2　充分性

统计量的充分性是统计中最重要的基本概念之一.它是费希尔于1922年提出来的,但充分性的严格叙述和证明是Halmos和Savage[8]在1949年完成的.我们将逐步深入讨论这个问题.

"不损失信息"的统计量就是充分统计量.这是一种模糊的说法.这里首先要对"不损失信息"给出明确的数学定义.设总体X服从某个离散分布$p_\theta(x)$,其中

$$p_\theta(x) = P_\theta(X = x), \quad x = a_1, a_2, \cdots$$

θ是未知参数(或向量),这里规定$p_\theta(a_i)>0, i=1,2,\cdots$,且$\sum_{i=1}^{\infty} p_\theta(a_i)=1$.为了对参数$\theta$作统计推断,需要从该总体抽取一个样本$X=(X_1,\cdots,X_n)$.显然,样本$X$中含有$\theta$的信息.经过加工,获得统计量$T=T(X)$,自然,统计量$T$也含有$\theta$的信息,试问:统计量$T$中和样本$X$中所含$\theta$的信息是否一样多?大家知道,对样本$X$加工不可能增加信息,不减少$\theta$的信息就是最好的了.由样本$X$可算出统计量$T$,假如能由统计量$T$的值恢复样本,那么这种统计量就不会损失有关$\theta$的信息,要做到这一点,关键要在给定$T=t$下,样本$X$的条件分布不依赖于$\theta$,即

$$P_\theta(X = x|T = t) = P(X = x|T = t) \tag{1.56}$$

为了说明这一点,我们先看下面的例子.

例 1.15　设$X_1,\cdots,X_n$是来自二点分布$b(1,\theta)$的一个样本,我们来研究如下二个统计量

$$T_1 = X_1 + \cdots + X_n, \quad T_2 = X_1X_2\cdots X_n.$$

显然,T_1的分布为二项分布$b(n,\theta)$,当T_1取$0,\cdots,n$中任一个整数t时,样本X的条件分布为

$$P_\theta(X = x|T_1 = t) = \frac{P_\theta(X = x, T_1 = t)}{P_\theta(T_1 = t)}$$

$$= \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \binom{n}{t}^{-1}, \quad \forall x$$

计算结果表明，条件分布 $P_\theta(X=x|T_1=t)$ 对任意样本点 x 都不依赖于 θ. 即此条件分布中已不含有 θ 的信息，样本中有关 θ 的信息全含在统计量 T_1 中.

另外，统计量 T_2 仅可能取两个值，0 与 1，且

$$P_\theta(T_2=1)=\theta^n,\quad P_\theta(T_2=0)=1-\theta^n$$

于是在 $T_2=0$ 下，样本 X 的条件分布为

$$\begin{aligned}&P_\theta(X_1=x_1,\cdots,X_n=x_n|T_2=0)\\=&P(X_1=x_1,\cdots,X_n=x_n,T_2=0)/P_\theta(T_2=0)\end{aligned}$$

这个条件概率尚不能具体得出，它还随着取值 $(x_1,\cdots,x_n)$ 而变. 譬如，当 $(x_1,\cdots,x_n)=(1,1,0,\cdots,0)$ 时，上述条件概率为 $\theta^2(1-\theta)^{n-2}/(1-\theta^n)$. 这表明在 $T_2=0$ 时，样本的条件分布依赖于 θ，若从此条件分布抽取样本，还可获得有关 θ 的信息，换句话说，T_2 没有包含样本中有关 θ 的全部信息，只包含部分，而损失部分.

进一步，我们可以利用 $T_1=t$，设计一个不依赖于 θ 的随机试验，使其产生与样本 X 有同样分布的新样本 $X'=(X'_1,\cdots,X'_n)$. 这是因为条件分布 $P(X_1=x_1,\cdots,X_n=x_n|T_1=t)=\binom{n}{t}^{-1}$ 已不依赖于 θ，且组合数 $\binom{n}{t}=n!/[t!(n-t)!]$ 可看作是 t 个 1 和 $n-t$ 个 0 的重复排列数. 因此可设计这样一个随机试验：把 t 个 1 和 $n-t$ 个 0 随机排列，任一个这样的排列出现都是等可能的. 若记 x'_i 为第 i 个位置上的数，则 x'_i 非 0 即 1，这样得到的 $(x'_1,\cdots,x'_n)$ 是新样本 $X'=(X'_1,\cdots,X'_n)$ 的观察值. 这样得到的新样本 X' 虽不能与老样本 X 完全相同，但它们在条件 $T_1=t$ 下的条件概率是相同的，都等于 $\binom{n}{t}^{-1}$. 故

$$\begin{aligned}&P_\theta(X_1=x_1,\cdots,X_n=x_n)\\=&\sum_{t=0}^{n}P(X_1=x_1,\cdots,X_n=x_n|T=t)P_\theta(T=t)\\=&\sum_{t=0}^{n}P(X'_1=x'_1,\cdots,X'_n=x'_n|T=t)P_\theta(T=t)\end{aligned}$$

$$= P_\theta(X'_1 = x'_1, \cdots, X'_n = x'_n)$$

由此推出新老样本是同分布的. 从而新老样本所含 θ 的信息是相同的，由于我们所设计的随机试验（古典概率模型）不含任何 θ 信息，所以老样本 X 中所含 θ 的信息全都在统计量 T_1 中. 能有如此结果关键在于条件分布 $P(X=x|T_1=t)$ 不依赖于 θ，而统计量 T_2 就不可能做到这一点，因为其条件分布 $P_\theta(X=x|T_2=t)$ 是依赖于 θ 的. 综合上述可知，T_1 是充分统计量，T_2 不是充分统计量.

从上述例子可以看出，在对样本的加工过程中，一个统计量 T"不损失信息"的数学描述是"在 T 取任一个值 t 时，样本的条件分布不依赖于未知参数". 当从离散分布族推广到一般分布族时，这一数学描述仍然有效，但允许 T 的一个零测集有例外. 由此可给出充分统计量的一般定义.

定义 1.18 设 $(\mathscr{X}, \mathscr{B}, \{P_\theta: \theta \in \Theta\})$ 是一个统计结构，又设 $T=T(X)$ 是 $(\mathscr{X}, \mathscr{B})$ 到 $(\mathscr{T}, \mathscr{C})$ 的一个统计量，P_θ^T 是 T 的诱导分布. 假如在 P_θ^T 的零测集外，T 取任一个值 t 时，样本 $X=(X_1, \cdots, X_n)$ 的条件分布都不依赖于 θ，即对任意的 $\theta \in \Theta$ 和 $B \in \mathscr{B}$，有

$$P_\theta(B|t) = P(B|t), \quad \text{a.s.}\ P_\theta^T \tag{1.57}$$

则称 T 为该分布族（或参数 θ）的充分统计量.

当样本空间 $\mathscr{X}$ 为欧氏空间，分布族 $\mathscr{P}=\{P_\theta: \theta \in \Theta\}$ 是受控的场合下，上述定义中的条件分布可用条件密度函数替代. 譬如在离散分布族时，T 是充分统计量的充要条件是

$$\begin{aligned} &P_\theta(X_1 = x_1, \cdots, X_n = x_n | T = t) \\ &= P(X_1 = x_1, \cdots, X_n = x_n | T = t) \end{aligned} \tag{1.58}$$

在连续分布族时，T 是充分统计量的充要条件是

$$p_\theta(x_1, \cdots, x_n | t) = p(x_1, \cdots, x_n | t), \quad \text{a.s.}\ P_\theta^T \tag{1.59}$$

其中 $p_\theta(x_1, \cdots, x_n|t)$ 是在给定 $T=t$ 下，$(X_1, \cdots, X_n)$ 的联合条件密度函数. 这两种场合是实际中最常遇到的场合.

在最一般的场合，确定条件分布并非易事，它需要从抽象的条件期望与条件概率的定义出发逐步地探讨. 这种探讨是冗长的和乏味的，有兴趣的读者可参阅陈希孺[1]和成平[2]的书. 在他们的书中，还在最一般

的场合讨论了充分统计量的一些性质.其中有两条性质在实际中常要用到,这里不加证明地引述如下.

定理 1.9 设$(\mathscr{X},\mathscr{B},\mathscr{P})$为统计结构,$\mathscr{X}$为欧氏空间,则次序统计量$(X_{(1)},\cdots,X_{(n)})$为该分布族$\mathscr{P}$的充分统计量.

这个结论与人们的如下经验事实相符.若n次试验都在相同条件下独立进行时,那我们只需知道n次试验的结果是什么,而试验结果的排列次序是无关紧要的,特别,当人们知道次序统计量的观察值时,并没有损失样本中任何有用的信息.

定理 1.10 充分统计量的一对一变换仍是充分统计量.

从直观上看来,任何一对一变换都不会损失信息,因为它能轻而易举地恢复原来充分统计量的取值.

在上面一些叙述的基础上我们来讨论一些分布族的充分统计量.

例 1.16 设$X_1,\cdots,X_n$是来自 Poisson 分布$P(\lambda)$的一个样本.证明:统计量$T=\sum_{i=1}^{n}X_i$是参数λ的充分统计量.

事实上,由 Poisson 分布的可加性知,$T\sim P(n\lambda)$,即

$$P(T=t)=\frac{(n\lambda)^t}{t!}\mathrm{e}^{-n\lambda},\quad t=0,1,\cdots$$

当$T=t$时,样本的条件分布为

$$\begin{aligned}
&P_\lambda(X_1=x_1,\cdots,X_n=x_n\mid T=t)\\
&=\frac{P(X_1=x_1)\cdots P(X_{n-1}=x_{n-1})P\left(X_n=t-\sum_{i=1}^{n-1}x_i\right)}{P(T=t)}\\
&=\left[\prod_{i=1}^{n-1}\frac{\lambda^{x_i}}{x_i!}\mathrm{e}^{-\lambda}\right]\frac{\lambda^{t-\sum_{i=1}^{n-1}x_i}\mathrm{e}^{-\lambda}}{\left(t-\sum_{i=1}^{n-1}x_i\right)!}\Bigg/\frac{(n\lambda)^t\mathrm{e}^{-n\lambda}}{t!}\\
&=\frac{t!}{x_1!\cdots x_{n-1}!\left(t-\sum_{i=1}^{n-1}x_i\right)!}\cdot\frac{1}{n^t}
\end{aligned}$$

这个条件分布是多项分布,它在t给定下就完全确定,它与参数λ无关,故T是参数λ的充分统计量.顺便指出,根据这个多项分布可以设计一个古典概率模型,把t个相同的球随机放入n个不同盒子,记X'_i为第i个盒子中球的个数,$i=1,\cdots,n$.则所得样本$(X'_1,\cdots,X'_n)$与原样

本同分布.

例 1.17　设 $X_1,\cdots,X_n$ 是来自两参数指数分布的一个样本,其密度函数为

$$p(x;\alpha,\beta)=\begin{cases}\frac{1}{\beta}\exp\left\{-\frac{x-\alpha}{\beta}\right\}, & x\geqslant\alpha\\ 0, & x\leqslant\alpha\end{cases}$$

其中 $\alpha\in\mathbf{R}$ 称为门限参数,$\beta\in\mathbf{R}^+$ 称为尺度参数.证明:$(X_{(1)},T^*_{n-1})=\left(X_{(1)},\sum_{i=2}^{n}X_{(i)}\right)$ 为 $\theta=(\alpha,\beta)$ 的充分统计量.其中 $X_{(1)},\cdots,X_{(n)}$ 为该样本的次序统计量.

我们将用一系列一对一变换来研究此问题,首先容易写出次序统计量 $(X_{(1)},\cdots,X_{(n)})$ 的联合密度函数

$$p_1(x_{(1)},\cdots,x_{(n)};\alpha,\beta)=\frac{n!}{\beta^n}\exp\left\{-\frac{1}{\beta}\left(\sum_{i=1}^{n}x_{(i)}-n\alpha\right)\right\},$$

$$\alpha\leqslant x_{(1)}\leqslant\cdots\leqslant x_{(n)}$$

作变换

$$\begin{cases}y_1=nx_{(1)}\\ y_2=(n-1)(x_{(2)}-x_{(1)})\\ \cdots\cdots\\ y_n=x_{(n)}-x_{(1)}\end{cases}\tag{1.60}$$

其雅可比行列式 $J=1/n!$.在此变换下,$\sum_{i=1}^{n}x_{(i)}=\sum_{i=1}^{n}y_i$,由此可得 $Y_1,\cdots,Y_n$ 的联合密度函数为

$$p_2(y_1,\cdots,y_n;\alpha,\beta)=\frac{1}{\beta^n}\exp\left\{-\frac{1}{\beta}\left(\sum_{i=1}^{n}y_i-n\alpha\right)\right\},$$

$$y_1\geqslant n\alpha,0\leqslant y_2,\cdots,y_n<\infty$$

再作变换

$$\begin{cases}y_1=y_1\\ U_1=y_2\\ U_2=y_2+y_3\\ \cdots\cdots\\ U_{n-2}=y_2+\cdots+y_{n-1}\\ T_{n-1}=y_2+\cdots+y_{n-1}+y_n\end{cases}\tag{1.61}$$

其雅可比行列式 $J_2=1$，在此变换下，$\sum_{i=1}^{n} y_i = y_1 + T_{n-1}$，由此可得 $Y_1, U_1, \cdots, U_{n-2}, T_{n-1}$ 的联合密度函数

$$p_3(y_1, u_1, \cdots, u_{n-2}, t) = \frac{1}{\beta^n}\exp\left\{-\frac{1}{\beta}(y_1 + t - n\alpha)\right\},$$

$$y_1 \geqslant n\alpha, 0 \leqslant u_1 \leqslant \cdots \leqslant u_{n-2} \leqslant t < \infty$$

因而 (Y_1, T_{n-1}) 的边际密度函数可求得

$$p_4(y_1, t) = \frac{1}{\beta^n}\exp\left\{-\frac{1}{\beta}(y_1 + t - n\alpha)\right\} \cdot$$

$$\int_0^t du_1 \int_{u_1}^t du_2 \cdots \int_{u_{n-3}}^t du_{n-2}$$

$$= \frac{1}{\beta^n}\exp\left\{-\frac{1}{\beta}(y_1 + t - n\alpha)\right\}\frac{t^{n-2}}{(n-2)!}$$

最后，在给定 $Y_1 = y_1$ 和 $T_{n-1} = t$ 时，$U_1, \cdots, U_{n-2}$ 的条件密度函数为

$$h(u_1, \cdots, u_{n-2} \mid y, t) = \frac{p_3(y, u_1, \cdots, u_{n-2}, t)}{p_4(y, t)} = \frac{(n-2)!}{t^{n-2}},$$

$$0 \leqslant u_1 \leqslant \cdots \leqslant u_{n-2} \leqslant t \qquad (1.62)$$

上式右端与 α, β 无关．所以 (Y_1, T_{n-1}) 是 (α, β) 的充分统计量．又因为

$$\begin{cases} Y_1 = nX_{(1)} \\ T_{n-1} = -(n-1)X_{(1)} + T_{n-1}^* \end{cases} \qquad (1.63)$$

是一对一变换，所以 $(X_{(1)}, T_{n-1}^*)$ 仍是 (α, β) 的充分统计量．

作为本例的结束．我们来讨论如何在充分统计量 $(X_{(1)}, T_{n-1}^*)$ 的给定下来恢复样本．设已知 $X_{(1)} = x_{(1)}, T_{n-1}^* = t^*$，则

(i) 由(1.63)算得 $Y_1 = nX_{(1)}, T_{n-1} = -(n-1)X_{(1)} + t^* = t$.

(ii) 由(1.62)看出，条件分布 h 是来自均匀分布 $U(0,t)$ 的容量为 $n-2$ 的样本的次序统计量的联合密度函数．故先从 $U(0,t)$ 随机抽取容量为 $n-2$ 的一个样本，然后从小到大排列即得 $u_1 \leqslant \cdots \leqslant u_{n-2}$.

(iii) 由(1.61)可从 $y_1, u_1, \cdots, u_{n-2}, t$ 算得 $y_1, \cdots, y_n$

(iv) 由(1.60)再算得 $x_{(1)} \leqslant \cdots \leqslant x_{(n)}$

(v) 把 $x_{(1)}, \cdots, x_{(n)}$ 随机排列就得到来自两参数指数分布的一个容量为 n 的样本．

§1.5.3　因子分解定理

验证一个统计量的充分性可以用定义 1.18. 在欧氏样本空间场合可以用条件分布，在可控结构场合（样本空间为欧氏的）可以用条件密度函数，除此之外，还有一个更方便判别充分统计量的方法，它就是因子分解定理. 这个判别法的简单情况由 Cramer 在[9]中证明，一般情况是 Halmos 和 Savage 在 1949 年证明的[8].

定理 1.11　（因子分解定理）设$(\mathscr{X},\mathscr{B},\{p_\theta:\theta\in\Theta\})$为可控结构，$\mu$为控制测度，记 $p_\theta(x)=\mathrm{d}P_\theta/\mathrm{d}\mu$，又设 $T=T(x)$是$(\mathscr{X},\mathscr{B})$到$(\mathscr{T},\mathscr{C})$上的统计量，则 T 为充分统计量的充要条件为存在

(i) $\mathscr{X}$ 上的非负 $\mathscr{B}$ 可测函数 $h(x)$，

(ii) $\mathscr{T}$ 上的 $\mathscr{C}$ 可测函数 $g_\theta(t)$，

使得对任意 $\theta\in\Theta$ 有

$$p_\theta(x)=g_\theta(T(x))h(x),\quad \text{a.s.}\,\mu \tag{1.64}$$

这个定理表明，在充分统计量存在的场合，样本的密度函数 $p_\theta(x)$ 可分解为两个因子的乘积，其中一个因子与 θ 无关，仅是样本 x 的函数，另一个因子是与 θ 有关，也与样本 x 有关，但与样本关系可通过充分统计量 $T(x)$表现出来. 反之，若样本的密度函数可写成(1.64)形式，其中$T(x)$一定是该分布的充分统计量，这无疑对寻找充分统计量是十分方便的.

证明：这个定理的一般证明可参阅文献[1],[2]，由于数学工具的限制，下面只给出离散场合下的证明. 这时

$$p_\theta(x)=P_\theta(X=x)=P_\theta(X_1=x_1,\cdots,X_n=x_n)$$

对任意固定的 $t\in\mathscr{T}$，其原像集合记为

$$A(t)=\{x:T(x)=t\}$$

必要性：设 $T(x)$是参数 θ 的充分统计量，则在给定 $T=t$ 下，条件概率 $P_\theta(X=x|T=t)$与参数 θ 无关，它只能是 x 的函数，这里记为 $h(x)$. 另外，无条件概率 $P_\theta(T=t)$可记为 $g_\theta(t)$. 于是对给定的 t 及 $x\in A(t)$，我们有

$$p_\theta(x)=P_\theta(X=x)$$

$$= P_\theta(X = x, T = t)$$
$$= P_\theta(X = x | T = t) P_\theta(T = t)$$
$$= h(x) \cdot g_\theta(t)$$

这就是因子分解形式(1.64).

充分性：设 $p_\theta(x)$ 有上述因子分解形式(1.64)，要证明条件概率 $P_\theta(X=x|T=t)$ 与 θ 无关，这只需对满足 $P_\theta(T=t)>0$ 的 t 证明即可. 分两种情况讨论，首先对任意 $x\in A(t)$ 可有

$$P_\theta(X = x | T = t) = \frac{P_\theta(X = x, T = t)}{P_\theta(T = t)}$$
$$= \frac{P_\theta(X = x)}{P_\theta(T = t)} = \frac{p_\theta(x)}{\sum\limits_{y\in A(t)} p_\theta(y)}$$

因为对 $y\in A(t)$，有 $T(y)=t$，故可用(1.64)代入上式，得

$$P_\theta(X = x | T = t) = \frac{g_\theta(t) h(x)}{\sum\limits_{y\in A(t)} g_\theta(t) h(y)} = \frac{h(x)}{\sum\limits_{y\in A(t)} h(y)}$$

最后的结果与参数 θ 无关. 其次，当 $x\notin A(t)$ 时，$T(x)\neq t$，于是事件"$X=x$"与事件"$T(x)=t$"不可能同时出现，所以当 $x\notin A(t)$ 时，$P_\theta(X=x, T=t)=0$，从而 $P_\theta(X=x|T=t)=0$，这也与参数 θ 无关. 综上所述，$T(x)$ 是充分统计量. 证毕.

下面的例子将告诉我们如何使用因子分解定理获得充分统计量. 当然，在连续分布场合，μ 取 Lebesgue 测度，在离散场合，μ 取计数测度，这些在实用场合都是默认的.

例 1.18 设 $X=(X_1,\cdots,X_n)$ 是来自 Poisson 分布 $P(\lambda)$ 的一个样本. 其样本的联合密度函数为

$$P(X_1 = x_1, \cdots, X_n = x_n) = \lambda^{\sum\limits_{i=1}^{n} x_i} e^{-n\lambda} \Big/ \prod_{i=1}^{n} (x_i!)$$

取 $T(x) = \sum\limits_{i=1}^{n} x_i$，$h(x)=\left(\prod\limits_{i=1}^{n}(x_i!)\right)^{-1}$，则上式可改写为

$$P(X = x) = [\lambda^{T(x)} e^{-n\lambda}] h(x)$$

由因子分解定理知，$T(x)=\sum_{i=1}^{n} x_i$ 是 λ 的充分统计量，这与例 1.16

所得结果是一致的. 可这里方便得多.

例 1.19 设 $X=(X_1,\cdots,X_n)$是来自均匀分布$U(0,\theta)$的一个样本. 其样本的联合密度函数为

$$p_\theta(x) = \theta^{-n} I_{(0,\theta)}(x_{(n)})$$

其中 $x_{(n)}$为最大次序统计量 $X_{(n)}$的取值. 由因子分解定理知, $X_{(n)}$是 θ 的充分统计量.

例 1.20 设 $X=(X_1,\cdots,X_n)$是来自正态分布 $N(\mu,\sigma^2)$的一个样本. 其样本的联合密度函数为

$$\begin{aligned} p_{\mu,\sigma^2}(x) &= (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu)^2\right\} \\ &= (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2\right]\right\} \end{aligned}$$

从因子分解定理角度去看, $\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$是$(\mu,\sigma^2)$的充分统计量. 但不可说 $\sum_{i=1}^n X_i$ 是 μ 的充分统计量, $\sum_{i=1}^n X_i^2$ 是 σ^2 的充分统计量, 因为在估计 σ^2 时, 仅用 $\sum_{i=1}^n X_i^2$ 是不够的, 在估计 μ^2 时, 仅用 $\sum_{i=1}^n X_i$ 也是不够的.

假如已知 μ, 则 $\sum_{i=1}^n (X_i-\mu)^2$ 是 σ^2 的充分统计量.

假如已知 σ^2, 则 $\sum_{i=1}^n X_i$ 或 $\overline{X}=\frac{1}{n}\sum_{i=1}^n X_i$ 是 μ 的充分统计量.

例 1.21 设 $X=(X_1,\cdots,X_n)$是来自两参数指数分布的一个样本, 其密度函数为

$$p_{\alpha,\beta}(x) = \begin{cases} \frac{1}{\beta}\exp\left(-\frac{x-\alpha}{\beta}\right), & x > \alpha \\ 0, & x \leqslant \alpha \end{cases}$$

记 $X_{(1)},\cdots,X_{(n)}$为该样本的次序统计量. 假如在实际中只观察到前 $r(\leqslant n)$个次序统计量的值后就停止试验, 这样的试验称为截尾试验, 它在产品的寿命试验中常见, 当 $r=n$ 时称为完全试验. 在截尾试验场合所得截尾样本$(X_{(1)},\cdots,X_{(r)})$的联合密度函数为

$$p_{\alpha,\beta}(x_{(1)},\cdots,x_{(r)}) = \frac{n!}{(n-r)!}\beta^{-r}\cdot$$

$$\exp\left\{-\frac{1}{\beta}\left[\sum_{i=1}^{r}x_i+(n-r)x_r-na\right]\right\}I_{(a,\infty)}(x_{(1)})$$

由因子分解定理知，$\left(X_{(1)},\sum_{i=1}^{r}X_{(i)}+(n-r)X_{(r)}\right)$为$(\alpha,\beta)$的充分统计量. 特别，当$r=n$时，$\left(X_{(1)},\sum_{i=1}^{n}X_{(i)}\right)$为$(\alpha,\beta)$的充分统计量. 这与例1.17所得充分统计量$\left(X_{(1)},\sum_{i=1}^{n}X_{(i)}\right)$是一致的. 因为它们之间都可通过一对一变换获得. 由定理1.10知，一对一变换是不会改变统计量的充分性.

§1.5.4 最小充分统计量

一个分布族$\mathscr{P}$的充分统计量往往不止一个，它们在“不损失信息”方面功能是相同的，为了取舍，我们就要比较它们在“压缩数据”的功能大小. 愈能压缩数据的充分统计量愈好. 而压缩数据功能的大小往往可通过其原像组成的σ代数的包含关系来判别. 下面的例子帮助我们更深入地探讨这个问题.

例1.22 设$(X_1,\cdots,X_n)$是来自二点分布$b(1,\theta)$的一个样本，其中$0<\theta<1$，则样本的联合分布列为

$$P(X_1=x_1,\cdots,X_n=x_n)=\theta^{\sum_{i=1}^{n}x_i}(1-\theta)^{n-\sum_{i=1}^{n}x_i}$$

其中$x_i=0,1,i=1,\cdots,n$. 根据因子分解定理，可以判定下面的n个统计量都是充分的.

$$\begin{aligned}
T_1&=(X_1,X_2,\cdots,X_n)\\
T_2&=(X_1+X_2,X_3,\cdots,X_n)\\
T_3&=(X_1+X_2+X_3,X_4,\cdots,X_n)\\
&\cdots\cdots\\
T_n&=X_1+X_2+\cdots+X_n.
\end{aligned}$$

为了比较这n个充分统计量的压缩数据的功能，我们来考察每个充分统计量的维数$\dim(T_k)$，T_k的取值空间$\mathscr{C}_k$中点的个数$N(T_k)$和T_k的一切原像组成的子σ代数$T_k^{-1}(\mathscr{C}_k)$所含事件的个数$\sigma(T_k)$，这些量的计算结果列于表1.2中.

表 1.2　几个充分统计量的比较

T_k	$\dim(T_k)$	$N(T_k)$	$\sigma(T_k)$	$n=10$ 时的 $\sigma(T_k)$
T_1	n	2^n	2^{2^n}	2^{1024}
T_2	$n-1$	$3\cdot 2^{n-2}$	$2^{3\cdot 2^{n-2}}$	2^{768}
T_3	$n-2$	$4\cdot 2^{n-3}$	$2^{4\cdot 2^{n-3}}$	2^{512}
⋮	⋮	⋮	⋮	⋮
T_n	1	$n+1$	2^{n+1}	2^{11}

从表中可以看出，无论在维数和其产生的子 σ 代数所含事件数等方面，统计量 T_n 是其中压缩数据功能最强的充分统计量. 现在问：是否还存在其它充分统计量，它比 T_n 更能压缩数据呢？回答是否定的. 事实上，倘若存在这样的统计量（记为 S），那它的维数不能比 T_n 的维数 $\dim(T_n)=1$ 更小了. 故只能在取值上比 T_n 更少一些，如今 T_n 可取 $n+1$ 个值. 我们定义 S 为一个仅取 n 个值的统计量

$$S(X_1,\cdots,X_n)=\begin{cases}1, & \text{当}\sum_{i=1}^{n}X_i=0\text{ 或 }1\\ k, & \text{当}\sum_{i=1}^{n}X_i=k,k=2,\cdots,n\end{cases}$$

显然 S 是 T_n 的函数，但 S 不是充分统计量，因为在 $S=1$ 时，条件概率

$$\begin{aligned}&P(X_1=0,\cdots,X_n=0\,|\,S=1)\\ =&\frac{P(X_1=0,\cdots,X_n=0)}{P(S=1)}\\ =&\frac{(1-\theta)^n}{(1-\theta)^n+n\theta(1-\theta)^{n-1}}=\frac{1-\theta}{1+(n-1)\theta}\end{aligned}$$

与参数 θ 有关，虽然这是一个特殊例子，但从中可见一般性，于是可得：T_n 是二点分布族中最能压缩数据的充分统计量，这种充分统计量称为最小充分统计量.

如何来识别最小充分统计量呢？为此我们研究两个统计量之间的关系. 设 T_i 是 $(\mathscr{X},\mathscr{B})$ 到 $(\mathscr{T}_i,\mathscr{C}_i)$ 上的统计量，$i=1,2$. 记 $\mathscr{B}_{0i}=T_i^{-1}(\mathscr{C}_i)$ 为由 T_i 产生的子 σ 代数. 若 $\mathscr{B}_{01}=\mathscr{B}_{02}$，则 T_1 与 T_2 在压缩数

据的功能相同，常称 T_1 与 T_2 等价，若 $\mathscr{B}_{01}\subset\mathscr{B}_{02}$，则在压缩数据的功能上 T_1 比 T_2 强. 利用可测函数性质[1]可知这时存在 $\mathscr{B}_{02}$ 可测函数 $f(T_2)$，使得 $T_1(x)=f(T_2(x))$，即 T_1 是 T_2 的函数，反之，若 T_1 是 T_2 的函数，则一定有 $\mathscr{B}_{01}\subset\mathscr{B}_{02}$. 由此可见，比较两个充分统计量在压缩数据功能上的大小，可观察其间的函数关系来确认. 当然不是任意两个充分统计量之间都有函数关系，假如存在这样的充分统计量，它是任一个其它充分统计量的函数，那么这个充分统计量是最能压缩数据，它们产生的 σ 代数也是最小的，这也就是最小充分统计量名称的由来. 它的一般定义如下.

定义 1.19 设 $T=T(x)$ 是定义在 $(\mathscr{X},\mathscr{B},\mathscr{P})$ 上的充分统计量，假如对 $(\mathscr{X},\mathscr{B},\mathscr{P})$ 上任一个充分统计量 $S=S(x)$，存在这样一个函数 $f(s)$，使得

$$T(x) = f(S(x)), \quad \forall\, x \in \mathscr{X} \tag{1.65}$$

则称 T 为 $(\mathscr{X},\mathscr{B},\mathscr{P})$ 的最小充分统计量.

在这个定义中，不同的 S 并不要求 f 是相同的，这个定义还可放宽要求，允许(1.65)在一个零测集不成立. 最小充分统计量并不永远存在. Bahaur 在 1954 证明，在因子分解定理成立的条件下，若样本空间为欧氏的，则最小充分统计量必存在. 常用的充分统计量(见前面例举的一些充分统计量)都是最小充分统计量. 它们常可用因子分解定理求出.

§1.6 完 备 性

§1.6.1 分布族的完备性

我们考虑一个参数统计结构 $(\mathscr{X},\mathscr{B},\mathscr{P})$，其中 $\mathscr{P}=\{p_\theta:\theta\in\Theta\}$，设 $\phi(x)$ 是定义在 $\mathscr{X}$ 上的 $\mathscr{B}$ 可测函数，一般说来，积分(如果存在的话)

$$\mathrm{E}_\theta\phi(x) = \int_{\mathscr{X}}\phi(x)\mathrm{d}P_\theta(x), \quad \theta \in \Theta$$

是参数 θ 的函数，因此，上述积分可以看作是一个变换，它把 $\mathscr{X}$ 上的一个函数 $\phi(x)$ 变换到 Θ 上的一个函数 $E_\theta\phi(x)$，在这个观点下，$E_\theta\phi(x)$ 可以看作是在这个积分下 $\phi(x)$ 的像.

这个变换在概率论与数理统计是要经常用到的，因为平均是消除随机性的好方法，有时我们还期望这个变换是一对一的变换，至少在几乎处处意义下是一对一变换，即

$$\phi_1(x) = \phi_2(x) \text{ a.s. } P_\theta \Leftrightarrow E_\theta\phi_1(x) = E_\theta\phi_2(x) \tag{1.66}$$

为此只需要"除了几乎处处为零的函数外，不存在其它函数使得它的像恒为零"，即

$$E_\theta\phi(x) = 0 \Rightarrow \phi(x) = 0, \quad \text{a.s. } P_\theta \tag{1.67}$$

因为，倘若 $\phi_1(x)$ 和 $\phi_2(x)$ 的像是相同的，那必导致

$$\int[\phi_1(x) - \phi_2(x)]\mathrm{d}P_\theta = 0$$

于是立即可从(1.67)推得(1.66)

$$\phi_1(x) = \phi_2(x), \quad \text{a.s. } P_\theta$$

而命题(1.67)成立与分布族 $\mathscr{P}$ 有极大的关系，不是任一个分布族都具有命题(1.67)这个性质，具有这个性质的分布族称为完备分布族.

定义 1.20　设 $(\mathscr{X},\mathscr{B},\mathscr{P})$ 是一个参数统计结构，其中 $\mathscr{P}=\{P_\theta: \theta\in\Theta\}$，假如对于 $\mathscr{B}$ 可测函数 $\phi(x)$，由

$$\int\phi(x)\mathrm{d}P_\theta = 0, \quad \forall\, \theta\in\Theta$$

总可推出 $\phi(x)=0$ a.s. P_θ，则称此结构 $(\mathscr{X},\mathscr{B},\mathscr{P})$ 为完备的，或称分布族 $\mathscr{P}$ 是完备的.

例 1.23　二项分布族是完备的.

二项分布族 $\mathscr{P}=\{p_p(x): 0<p<1\}$，其中 $p_p(x)=\binom{n}{x}p^x(1-p)^{n-x}$，$x=0,1,2,\cdots,n$，假如 $\phi(x)$ 满足

$$E_p\phi(x) = \sum_{x=0}^{n}\phi(x)\binom{n}{x}p^x(1-p)^{n-x} = 0$$

由于 $(1-p)^{n-x} = \sum_{k=0}^{n-x}(-p)^k\binom{n-x}{k}$，所以

$$\sum_{x=0}^{n}\sum_{k=0}^{n-x}\binom{n}{x}\binom{n-x}{k}(-1)^k\phi(x)p^{x+k}=0$$

上式左边是 p 的多项式,它对任意 $p\in(0,1)$均为零,这只有其系数几乎处处均为零才能做到,即 $\phi(x)=0$,a.s. P_p.所以二项分布族是完备的.

例 1.24 正态分布族$\{N(\mu,1),\mu\in\mathbf{R}\}$是完备的.

如果 $\phi(x)$满足 $\mathrm{E}_\mu\phi(x)=0,\quad \forall\ \mu\in\mathbf{R}$,即

$$\int_{-\infty}^{\infty}\phi(x)\mathrm{e}^{-\frac{1}{2}(x-\mu)^2}\mathrm{d}x=0,\quad \forall\ \mu\in\mathbf{R}$$

$$\int_{-\infty}^{\infty}\phi(x)\mathrm{e}^{-\frac{1}{2}x^2}\cdot\mathrm{e}^{\mu x}\mathrm{d}x=0,\quad \forall\ \mu\in\mathbf{R}$$

上式左边是函数 $\phi(x)\mathrm{e}^{-\frac{1}{2}x^2}$的双边拉氏变换,根据拉氏变换的唯一性知

$$\phi(x)\mathrm{e}^{-\frac{1}{2}x^2}=0,\quad \text{a.s. } P_\mu$$

由于 $\mathrm{e}^{-\frac{1}{2}x^2}\neq 0$,所以

$$\phi(x)=0,\quad \text{a.s. } P_\mu$$

这说明正态分布族$\{N(\mu,1):\mu\in\mathbf{R}\}$是完备的.

不完备的分布族也是常见的.

例 1.25 正态分布族$\{N(0,\sigma^2):\sigma\in\mathbf{R}^+\}$是不完备的.

要说明一个分布族是不完备的,只要能找到这样一个函数 $\phi(x)$它能使是 $\mathrm{E}_\theta\phi(x)=0$,但 $\phi(x)$本身不是几乎处处为零即可.

因为这个分布族中任一个密度函数都是偶函数,故对 $\phi(x)=x$(其它奇函数都可以)就有

$$\mathrm{E}_\sigma\phi(x)=0,\quad \forall\ \sigma\in\mathbf{R}^+$$

但这个 x 是服从正态分布 $N(0,\sigma^2)$的随机变量 X 取值,显然有

$$P(X\neq 0)=1$$

所以这个正态分布族$\{N(0,\sigma^2):\sigma\in\mathbf{R}^+\}$是不完备的.

§1.6.2 完备统计量

定义 1.21 设 T 是统计结构$(\mathscr{X},\mathscr{B},\mathscr{P})$上的统计量.$(\mathscr{T},\mathscr{C},$

$\mathscr{P}^T$)是由 T 诱导的统计结构，假如后者是完备结构，则称 T 为完备统计量.

显然，假如结构($\mathscr{X}$，$\mathscr{B}$，$\mathscr{P}$)是完备的，则统计量 T 的诱导结构($\mathscr{T}$，$\mathscr{C}$，$\mathscr{P}^T$)也是完备的，这是因为，若 $\phi(t)$满足

$$\mathrm{E}_\theta\phi(t) = 0, \quad \forall\, \theta \in \Theta$$

则由积分变换可知

$$\mathrm{E}_\theta\phi(t) = \int_{\mathscr{T}} \phi(t)\mathrm{d}P_\theta^T = \int_{\mathscr{X}} \phi(t(x))\mathrm{d}P_\theta(x)$$

所以由原结构的完备性可得

$$\phi(t(x)) = 0, \quad \text{a.s. } P_\theta$$

从而

$$\phi(t) = 0, \quad \text{a.s. } P_\theta^T$$

但是可能会出现这种情况，诱导结构是完备的，而原结构是不完备的.

例 1.26 在例 1.25 中已证明了正态分布$\{N(0,\sigma^2):\sigma\in\mathbf{R}^+\}$是不完备的，但是，统计量 $T_n = \sum_{i=1}^n X_i^2$ 都是完备统计量，其中$(X_1,\cdots,X_n)$是取自正态总体 $N(0,\sigma^2)$的一个子样.

事实上，$T_n = \sum_{i=1}^n X_i^2 \sim Ga\left(\dfrac{n}{2},\dfrac{1}{2\sigma^2}\right)$，其密度函数为

$$p_\sigma(t) = \frac{1}{2^{\frac{n}{2}}\sigma^n\Gamma(\frac{n}{2})}t^{\frac{n}{2}-1}\mathrm{e}^{-\frac{t}{2\sigma^2}}, \qquad t > 0$$

假如 $\phi(t)$满足 $\mathrm{E}_\sigma\phi(t)=0$，$\forall\,\sigma\in\mathbf{R}^+$，即

$$\int_0^\infty \phi(t)t^{\frac{n}{2}-1}\mathrm{e}^{-\frac{t}{2\sigma^2}}\mathrm{d}t = 0, \quad \forall\,\sigma \in \mathbf{R}^+$$

上式左边是函数 $\phi(t)t^{\frac{n}{2}-1}$的单边拉普拉斯变换，由拉氏变换的唯一性知

$$\phi(t)t^{\frac{n}{2}-1} = 0, \quad \text{a.s. } P_\sigma^T$$

当 $t>0$ 时，$t^{\frac{n}{2}-1}$不恒为零，所以

$$\phi(t) = 0, \quad \text{a.s. } P_\sigma^T$$

从而诱导分布族$\{p_\sigma(t):\sigma\in\mathbf{R}^+\}$是完备的，统计量$T_n=\sum_{i=1}^n X_i^2$也是完备的.

§1.7 指数结构

§1.7.1 定义与例子

定义 1.22 设$(\mathscr{X},\mathscr{B},\{p_\theta:\theta\in\Theta\})$是可控参数统计结构，假如其密度函数可表示为如下形式

$$p_\theta(x)=c(\theta)\exp\Big\{\sum_{j=1}^k c_j(\theta)T_j(x)\Big\}h(x) \tag{1.68}$$

并且它的支撑$\{x:p_\theta(x)>0\}$不依赖于θ，则称此结构为指数型的统计结构，简称指数结构，其中的分布族称为指数族，这里的$0<c(\theta)$，$c_1(\theta),\cdots,c_k(\theta)<\infty$，诸$T_j(x)$都是与$\theta$无关，且取有限值的$\mathscr{B}$可测函数，$k$为正整数，$h(x)>0$.

指数结构在数理统计中发挥着重要作用，理论上的许多重要问题只在指数结构中才有比较彻底的解决. 指数型分布族概括了大量的常用分布族的共同特征.

例 1.27 正态分布族$\{N(\mu,\sigma^2):\theta=(\mu,\sigma^2)\in\mathbf{R}\times\mathbf{R}^+\}$是指数型分布族，因为它对 Lebesgue 测度的密度函数为

$$p_\theta(x)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}=\frac{1}{\sqrt{2\pi}\sigma}\exp\Big\{-\frac{\mu^2}{2\sigma^2}-\frac{x^2}{2\sigma^2}+\frac{\mu}{2\sigma^2}x\Big\}$$

取$c(\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{\mu^2}{2\sigma^2}}$，$c_1(\mu,\sigma)=\frac{\mu}{\sigma^2}$，$c_2(\mu,\sigma)=-\frac{1}{2\sigma^2}$，$h(x)=1$ 于是

$$p_\theta(x)=c(\mu,\sigma)\exp\{c_1(\mu,\sigma)x+c_2(\mu,\sigma)x^2\}$$

特别，$X=(X_1,\cdots,X_n)$是来自正态分布$N(\mu,\sigma^2)$的一个样本时，其样本的联合密度函数可表示为

$$p_\theta(x)=[c(\mu,\sigma)]^n\exp\Big\{c_1(\mu,\sigma)\sum_{i=1}^n x_i+c_2(\mu,\sigma)\sum_{i=1}^n x_i^2\Big\}$$

它仍然是指数型分布族.

例 1.28 二项分布族是指数分布族.

因为它对计数测度的概率密度函数为

$$
\begin{aligned}
p_\theta(x) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \\
&= \binom{n}{x}(1-\theta)^n e^{x\ln\frac{\theta}{1-\theta}} \\
&= c(\theta)\exp\{c_1(\theta)x\}h(x), \quad x = 0,1,\cdots,n
\end{aligned}
$$

其中 $c(\theta)=(1-\theta)^n, c_1(\theta)=\ln\dfrac{\theta}{1-\theta}, h(x)=\dbinom{n}{x}$.

例 1.29 均匀分布族 $\{R(0,\theta):\theta\in\mathbf{R}^+\}$ 不是指数型分布族,因为它的支撑 $\{x:p_\theta(x)>0\}$ 依赖于未知参数 θ,同样,双参数指数分布族

$$p_{(\mu,\sigma)}(x) = \frac{1}{\sigma}e^{-\frac{x-\mu}{\sigma}}, \quad x\geqslant\mu,(\mu,\sigma)\in\mathbf{R}\times\mathbf{R}^+$$

的支撑 $\{x:p_{\mu,\sigma}(x)>0\}=(\mu,\infty)$ 依赖于未知参数 μ,所以它不是指数型分布族.假如 μ 已知,那它就是指数型分布族,因为此时,其支撑不依赖于未知参数 σ,而其密度函数又可写为

$$p_\sigma(x) = c(\sigma)\exp\{c_1(\sigma)(x-\mu)\}h(x)$$

其中 $c(\sigma)=\dfrac{1}{\sigma}, c_1(\sigma)=-\dfrac{1}{\sigma}, h(x)=1$ 又如 Weibull 分布族

$$p_{\eta,m}(x) = \frac{m}{\eta}\left(\frac{x}{\eta}\right)^{m-1}\exp\left\{-\left(\frac{x}{\eta}\right)^m\right\}, x>0,(\eta,m)\in\mathbf{R}^+\times\mathbf{R}^+$$

不是指数型分布族,因为在 m 未知时,x^m 不能用有限个 x 的函数表示.假如 m 已知,它是指数型分布族.

例 1.30 Gamma 分布族 $\{Ga(\alpha,\lambda):\alpha\in\mathbf{R}^+,\lambda\in\mathbf{R}^+\}$ 是指数型分布族,因为它对 Lebesgue 测度的密度函数为

$$
\begin{aligned}
p_{\alpha,\lambda}(x) &= \frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x} \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)}\exp\{-\lambda x+(\alpha-1)\ln x\} \\
&= c(\alpha,\lambda)\exp\{c_1(\alpha,\lambda)x+c_2(\alpha,\lambda)\ln x\}, \quad x>0
\end{aligned}
$$

其中 $c(\alpha,\lambda)=\lambda^{\alpha}/\Gamma(\alpha)$，$c_1(\alpha,\lambda)=-\lambda$，$c_2(\alpha,\lambda)=(\alpha-1)$.

假如在Gamma分布中引入第三个参数——门限参数 μ，其密度函数为

$$p_{\mu,\alpha,\lambda}(x)=\frac{\lambda^{\alpha}}{\Gamma(\alpha)}(x-\mu)^{\alpha-1}e^{-\lambda(x-\mu)},\quad x>\mu$$

这种三参数Gamma分布族就不是指数型分布族，因为其支撑依赖于未知的门限参数.

§1.7.2 指数型分布族的标准形式

若在指数型分布(1.68)中重新设置新参数 $\boldsymbol{w}=(w_1,\cdots,w_k)$，其中

$$w_j=c_j(\theta),\quad j=1,\cdots,k$$

假如能从此函数方程组中唯一地解出 $\theta=\theta(w_1,\cdots,w_k)$，那么再令 $c^*(\boldsymbol{w})=c(\theta(\boldsymbol{w}))$. 把这些代回(1.68)，即得密度函数的另一形式

$$p_{w}(x)=c^*(\boldsymbol{w})\exp\left\{\sum_{j=1}^{k}w_jT_j(x)\right\}h(x) \tag{1.69}$$

(1.69)与(1.68)在函数结构上并无本质变化，它们主要差别在参数选择上，因此老参数 θ 的参数空间 Θ 与新参数 $\boldsymbol{w}$ 的参数空间(记为 Ω)会有所变化，形如(1.68)的分布是先在 Θ 中选出 θ、然后通过计算 $c(\theta)$，$c_1(\theta)$，…，$c_k(\theta)$才能确定下来. 而形如(1.69)的分布，只要在 Ω 中选出 $\boldsymbol{w}$，计算 $c^*(\boldsymbol{w})$就可确定. 在研究中形式(1.69)特别方便，故特冠取"标准形式"名称.

定义 1.23 形如(1.69)的密度函数称为指数型分布族(1.68)的标准形式. 其中新参数 $\boldsymbol{w}$ 使在 $\mathscr{X}$ 上的积分

$$0<\int_{\mathscr{X}}\exp\left\{\sum_{j=1}^{k}w_jT_j(x)\right\}h(x)\mathrm{d}\mu(x)<\infty$$

成立的空间称为自然参数空间，记为 Ω，即

$$\Omega=\left\{\boldsymbol{w}=(w_1,\cdots,w_k):0<\int_{\mathscr{X}}\exp\left\{\sum_{j=1}^{k}w_jT_j(x)\right\}h(x)\mathrm{d}\mu(x)<\infty\right\} \tag{1.70}$$

例 1.31 二项分布族是指数分布族，它的密度函数为

$$p_\theta(x)=(1-\theta)^n e^{x\ln\frac{\theta}{1-\theta}}\binom{n}{x},x=0,\cdots,n,\theta\in(0,1)$$

若令

$$w=\ln\frac{\theta}{1-\theta}$$

可以解出

$$\theta=\frac{e^w}{1+e^w}$$

所以二项分布族的标准形式为

$$p_w(x)=(1+e^w)^{-n}e^{wx}\binom{n}{x},\qquad x=0,\cdots,n$$

其自然参数空间 $\Omega=(-\infty,\infty)$.

例 1.32　正态分布族是指数型分布族，它的密度函数为

$$p_{\mu,\sigma}(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{\mu^2}{2\sigma^2}}\exp\left\{\frac{\mu}{\sigma^2}x-\frac{1}{2\sigma^2}x^2\right\},$$

$$(\mu,\sigma)=\mathbf{R}\times\mathbf{R}^{+}$$

若令

$$w_1=\frac{\mu}{\sigma^2},\quad w_2=-\frac{1}{2\sigma^2}$$

可以解出

$$\mu=-\frac{w_1}{2w_2},\quad \sigma=\sqrt{-\frac{1}{2w_2}}$$

所以正态分布族的标准形式为

$$p_{w_1,w_2}(x)=\sqrt{\frac{-w_2}{n}}e^{-\frac{w_1^2}{w_2}}\exp\{w_1x+w_2x^2\}$$

其自然参数空间 $\Omega=\mathbf{R}\times\mathbf{R}^{-}$.

§1.7.3　指数型分布族的基本性质

下面用指数型分布族的标准形式(1.69)讨论其基本性质.

定理 1.12　自然参数空间 Ω 为凸集.

证明:因为

$$\Omega=\left\{\boldsymbol{w}=(w_1,\cdots,w_k):\int_{\mathscr{X}}\exp\left\{\sum_{j=1}^{k}w_jT_j(x)\right\}\mathrm{d}\mu_1(x)<\infty\right\}$$

其中 $\mathrm{d}\mu_1(x)=h(x)\mathrm{d}\mu(x)$,设 $\boldsymbol{w}'=(w'_1,\cdots,w'_k)$,$\boldsymbol{w}''=(w''_1,\cdots,w''_k)\in\boldsymbol{\Omega}$,又设 $0<\alpha<1$,往证 $\boldsymbol{w}=\alpha\boldsymbol{w}'+(1-\alpha)\boldsymbol{w}''=(w_1,\cdots,w_k)\in\boldsymbol{\Omega}$. 事实上

$$\int_{\mathscr{X}}\exp\left\{\sum_{j=1}^{k}w_jT_j(x)\right\}\mathrm{d}\mu_1(x)$$

$$=\int_{\mathscr{X}}\left[\exp\left\{\sum_{j=1}^{k}w'_jT_j(x)\right\}\right]^{\alpha}\left[\exp\left\{\sum_{j=1}^{k}w''_jT_j(x)\right\}\right]^{1-\alpha}\mathrm{d}\mu_1(x)$$

$$\leqslant\left[\int_{\mathscr{X}}\exp\left\{\sum_{j=1}^{k}w'_jT_j(x)\right\}\mathrm{d}\mu_1(x)\right]^{\alpha}\left[\int_{\mathscr{X}}\exp\left\{\sum_{j=1}^{k}w''_jT_j(x)\right\}\mathrm{d}\mu_1(x)\right]^{1-\alpha}$$

最后一个不等式是采用了 Hölder 不等式,由于 $\boldsymbol{w}',\boldsymbol{w}''\in\boldsymbol{\Omega}$,所以不等式右边两个因子中的积分存在,从而得到 $\boldsymbol{w}\in\Omega$. 这说明了 Ω 是凸集,证毕.

这定理告知,若 Ω 在 $\mathbf{R}^k$ 中有内点,则 Ω 的全部内点构成 $\mathbf{R}^k$ 上的一个凸区域,特别在 $k=1$ 场合,Ω 为一区间(有限或无限,开的,闭的或半开半闭的都有可能).

定理 1.13 设 $\phi(x)$是 $\mathscr{X}$ 上的 $\mathscr{B}$ 可测函数,且对一切 $\boldsymbol{w}=(w_1,\cdots,w_k)\in\boldsymbol{\Omega}$,有(以下把 $h(x)$吸收到 $\mathrm{d}\mu$ 中去,仍记为 $\mathrm{d}\mu$)

$$\int|\phi(x)|\exp\left\{\sum_{j=1}^{k}w_jT_j(x)\right\}\mathrm{d}\mu<\infty \tag{1.71}$$

若记 $z_j=w_j+\mathrm{i}\eta_j$,$j=1,\cdots,k$,即将 w_j 看作某复变量的实部,则

(i)积分

$$A(\boldsymbol{z})=\int_{\mathscr{X}}\phi(x)\exp\left\{\sum_{j=1}^{k}z_jT_j(x)\right\}\mathrm{d}\mu \tag{1.72}$$

在区域

$$A=\{z=(z_1,\cdots,z_k),z_j=w_j+\mathrm{i}\eta_j,j=1,\cdots,k,(w_1,\cdots,w_k)\in\boldsymbol{\Omega}\}$$

内每点有意义,且为 k 个复变量 $z_1,\cdots,z_k$ 的解析函数.

(ii)在区域 A 内,$A(\boldsymbol{z})$对每个 z_j 的任意阶偏导数可通过(1.72)式在积分号下求导数获得,即对(1.72)式积分和求导运算可以交换.

证明：由于$(w_1,\cdots,w_k)\in\Omega$和

$$\left|\phi(x)\exp\left\{\sum_{j=1}^{k}(w_j+\mathrm{i}\eta_j)T_j(x)\right\}\right|$$

$$=|\phi(x)|\exp\left\{\sum_{j=1}^{k}w_jT_j(x)\right\}$$

所以积分(1.71)在A内每一点有意义，这里使用了

$$|\exp\{\mathrm{i}\eta_jT_j(x)\}|=1$$

以下把(i)的后半部分和(ii)结合起来证明．现任取A的一内点$\mathbf{z}^0=(z_1^0,\cdots,z_k^0)=(w_1^0+\mathrm{i}\eta_1^0,\cdots,w_k^0+\mathrm{i}\eta_k^0)$，这意味着$(w_1^0,\cdots,w_k^0)$亦为$\Omega$的内点．现考察$A(\mathbf{z})$对$z_1$在$z_1^0$处的导数，在这个过程中，就把$z_2,\cdots,z_k$看作固定的．为此把$\phi(x)\exp\left\{\sum_{j=2}^{k}z_jT_j(x)\right\}$分解为实部和虚部，而每一个又可分解为正部与负部，即

$$\phi(x)\mathrm{e}^{\sum_{j=2}^{k}w_jT_j}\cdot\mathrm{e}^{\mathrm{i}\sum_{j=2}^{k}\eta_jT_j}$$

$$=\phi(x)\mathrm{e}^{\sum_{j=2}^{k}w_jT_j}\cos\left(\sum_{j=2}^{k}\eta_jT_j\right)+\mathrm{i}\phi(x)\mathrm{e}^{\sum_{j=2}^{k}w_jT_j}\sin\left(\sum_{j=2}^{k}\eta_jT_j\right)$$

$$=A_1(x)+\mathrm{i}A_2(x)$$

$$=A_1^+(x)-A_1^-(x)+\mathrm{i}A_2^+(x)-\mathrm{i}A_2^-(x)$$

其中

$$A_r^+(x)=\begin{cases}0, & 当\ A_r(x)<0\\ A_r(x), & 当\ A_r(x)\geqslant 0\end{cases}\qquad r=1,2$$

$$A_r^-(x)=\begin{cases}0, & 当\ A_r(x)>0\\ -A_r(x), & 当\ A_r(x)\leqslant 0\end{cases}$$

再引进如下测度

$$\mathrm{d}\mu_1=A_1^+(x)\mathrm{d}\mu,\ \mathrm{d}\mu_2=A_1^-(x)\mathrm{d}\mu$$

$$\mathrm{d}\mu_3=A_2^+(x)\mathrm{d}\mu,\ \mathrm{d}\mu_4=A_2^-(x)\mathrm{d}\mu$$

于是$A(\mathbf{z})$作为z_1的函数可以写成

$$A(\mathbf{z})=\int\mathrm{e}^{z_1T_1}\mathrm{d}\mu_1-\int\mathrm{e}^{z_1T_1}\mathrm{d}\mu_2+\mathrm{i}\int\mathrm{e}^{z_1T_1}\mathrm{d}\mu_3-\mathrm{i}\int\mathrm{e}^{z_1T_1}\mathrm{d}\mu_4$$

由于上式右端四个积分形式相同，所以只要能证明第一个积分为解析

函数就可以了,其余全可以类似证得.

因为$(w_1^0,\cdots,w_k^0)$是Ω的内点,故存在$\delta>0$,使得当$|w_1-w_1^0|<\delta$时,$(w_1+i\eta_1^0,z_2^0,\cdots,z_k^0)$是$A$的内点,从而使

$$h(z_1)=\int_{\mathscr{X}}e^{z_1T_1(x)}d\mu_1(x) \tag{1.73}$$

存在且有限,于是,差商的模

$$\left|\frac{h(z_1)-h(z_1^0)}{z_1-z_1^0}\right|\leqslant\int_{\mathscr{X}}\left|e^{z_1^0T_1}\frac{e^{(z_1-z_1^0)T_1}-1}{z_1-z_1^0}\right|d\mu_1$$

由于对任意实数a和复数z,在$|z|<\delta$时,有

$$\left|\frac{e^{az}-1}{z}\right|\leqslant\sum_{n=1}^{\infty}\frac{|a|^n}{n!}|z|^{n-1}\leqslant\sum_{n=1}^{\infty}\frac{|a|^n}{n!}\delta^{n-1}\leqslant\frac{e^{|a|\delta}}{\delta}$$

故知,在$|w_1-w_1^0|=|z_1-z_1^0|<\delta$内,上述积分中的被积函数不超过

$$\left|e^{z_1^0T_1}\right|\frac{1}{\delta}e^{\delta|T_1|}=\frac{1}{\delta}\left|e^{z_1^0T_1+\delta|T_1|}\right|$$

但是

$$z_1^0T_1+\delta|T_1|=\begin{cases}(z_1^0+\delta)T_1, & \text{当 } T_1>0\\(z_1^0-\delta)T_1, & \text{当 } T_1<0\end{cases}$$

于是

$$\left|e^{z_1^0T_1}\frac{e^{(z_1-z_1^0)T_1}-1}{z_1-z_1^0}\right|\leqslant\frac{1}{\delta}\left|e^{(z_1^0+\delta)T_1}\right|+\frac{1}{\delta}\left|e^{(z_1^0-\delta)T_1}\right|$$

由于$(z_1^0\pm\delta,z_2^0,\cdots,z_k^0)\in A$,故上式右端可积,从而对(1.73)式可使用Lebesgue控制收敛定理,极限号与积分号交换,于是得

$$\begin{aligned}\lim_{z_1\to z_1^0}\frac{h(z_1)-h(z_1^0)}{z_1-z_1^0}&=\int_{\mathscr{X}}\lim_{z_1\to z_1^0}\left[e^{z_1^0T_1}\frac{e^{(z_1-z_1^0)T_1}-1}{z_1-z_1^0}\right]d\mu_1\\&=\int_{\mathscr{X}}e^{z_1^0T_1(x)}T_1(x)d\mu_1(x)\end{aligned}$$

上式右端正是$h(Z)$对z_1在z_1^0处积分号下求导的结果.

使用类似方法,并利用归纳法,即可证明,$A(z)$对$z_1,\cdots,z_k$的各阶导数存在,且可在积分号下求导,证毕.

定理1.14 设$X=(X_1,\cdots,X_n)$是来自指数型分布标准形式

(1.69)的一个样本，则有

(i)统计量

$$(T_1(X),\cdots,T_k(X))=\left(\sum_{i=1}^{n}T_1(x_i),\cdots,\sum_{i=1}^{n}T_k(x_i)\right) \quad (1.74)$$

是指数型分布族的充分统计量.

(ii)充分统计量的期望和协方差分别为

$$\mathrm{E}_w(T_j(X))=-\frac{\partial \ln c(\boldsymbol{w})}{\partial w_j},\quad j=1,\cdots,k \quad (1.75)$$

$$\mathrm{Cov}_w(T_i(X),T_j(X))=-\frac{\partial^2 \ln c(\boldsymbol{w})}{\partial w_i \partial w_j},\quad i,j=1,\cdots,k \quad (1.76)$$

其中 $c(\boldsymbol{w})=[c^*(\boldsymbol{w})]^n$.

证明：由(1.69)可写出样本 X 的联合密度函数

$$p_w(X)=c(\boldsymbol{w})\exp\left\{\sum_{j=1}^{k}w_jT_j(X)\right\}h(X) \quad (1.77)$$

其中 $c(\boldsymbol{w})=[c^*(\boldsymbol{w})]^n$，$h(X)=h(x_1)\cdots h(x_n)$，诸 $T_j(X)$ 如(1.74)所示.由因子分解定理可得(i).

以下再证(ii).(1.77)对控制测度 $\mathrm{d}\mu$ 在 $\mathscr{X}$ 上的积分为 1，即

$$c(\boldsymbol{w})\int_{\mathscr{X}}\exp\left\{\sum_{j=1}^{k}w_jT_j(X)\right\}h(X)\mathrm{d}\mu=1$$

由于(1.77)仍为指数型分布，故可对上式两边求导，得

$$\frac{\partial c(\boldsymbol{w})}{\partial w_j}\int_{\mathscr{X}}\exp\left\{\sum_{j=1}^{k}w_jT_j(X)\right\}h(X)\mathrm{d}\mu+$$

$$c(\boldsymbol{w})\int_{\mathscr{X}}T_j\left(\exp\left\{\sum_{j=1}^{k}w_jT_j(X)\right\}h(X)\mathrm{d}\mu\right)=0$$

$$\frac{\partial c(\boldsymbol{w})}{\partial w_j}\cdot\frac{1}{c(w)}+\mathrm{E}_w[T_j(X)]=0$$

由此可得(1.75).再对(1.75)两边求导，可得

$$\frac{\partial}{\partial w_i}\left(\frac{\partial \ln c(\boldsymbol{w})}{\partial w_j}\right)=\frac{\partial}{\partial w_i}(-\mathrm{E}_w[T_j(X)])$$

$$=\frac{\partial}{\partial w_i}\left(-c(\boldsymbol{w})\int_{\mathscr{X}}T_j(x)\exp\left\{\sum_{j=1}^{k}w_jT_j(x)\right\}h(x)\mathrm{d}\mu\right)$$

$$= -\frac{\partial c(\boldsymbol{w})}{\partial w_i} \cdot \frac{\mathrm{E}_w[T_j(X)]}{c(\boldsymbol{w})} -$$

$$c(\boldsymbol{w}) \int_{\mathscr{X}} T_i(x)T_j(x)\exp\left\{\sum_{j=1}^{k} w_j T_j(x)\right\} h(x)\mathrm{d}\mu$$

$$= \mathrm{E}_w[T_i(X)] \cdot \mathrm{E}_w[T_j(X)] - \mathrm{E}_w[T_i(X)T_j(X)]$$

$$= -\mathrm{Cov}_w(T_i(X), T_j(X))$$

这就证明了(1.75)式,证毕.

例 1.33 设 X 服从对数正态分布 $LN(\mu,\sigma^2)$,当 $x>0$ 时,对数正态分布的密度函数为

$$p_{\mu,\sigma^2} = \frac{1}{\sqrt{2\pi}\sigma x}\exp\left\{-\frac{(\ln x-\mu)^2}{2\sigma^2}\right\}$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{\mu^2}{2\sigma^2}}\exp\left\{\frac{\mu}{\sigma^2}\ln x - \frac{1}{2\sigma^2}(\ln x)^2\right\} \cdot \frac{1}{x}$$

若令 $w_1=\mu/\sigma^2, w_2=-1/2\sigma^2$,则其标准形式为

$$p_{\mu,\sigma^2}(x) = c^*(w_1,w_2)\exp\{w_1\ln x + w_2(\ln x)^2\}h(x)$$

其中

$$c^*(w_1,w_2) = \sqrt{-\frac{w_2}{\pi}}\exp\left\{\frac{w_1^2}{4w_2}\right\}, \quad h(x) = 1/x$$

若从该分布抽取一个容量为 n 的样本$(X_1,\cdots,X_n)$.则容易看出,其充分统计量为

$$(T_1(X),T_2(X)) = \left(\sum_{i=1}^{n}\ln X_i, \sum_{i=1}^{n}(\ln X_i)^2\right)$$

以下来求它们的一,二阶矩,为此先计算 $c(w_1,w_2)=[c^*(w_1,w_2)]^n$ 的对数.然后用(1.75)和(1.76)求出它们的期望.方差和协方差.

$$\ln c(w_1,w_2) = \frac{n}{2}\ln\left(-\frac{w_2}{\pi}\right) + \frac{nw_1^2}{4w_2}$$

$$\frac{\partial \ln c}{\partial w_1} = \frac{nw_1}{2w_2} = -n\mu, \text{故 } \mathrm{E}[T_1(X)] = n\mu$$

$$\frac{\partial^2 \ln c}{\partial w_1^2} = \frac{n}{2w_2} = -n\sigma^2, \text{故 } \mathrm{Var}[T_1(X)] = n\sigma^2$$

$$\frac{\partial \ln c}{\partial w_2} = \frac{n}{2w_2} - \frac{nw_1^2}{4w_2^2} = -n(\sigma^2 + \mu^2),\text{故 } \mathrm{E}[T_2(X)] = n(\sigma^2 + \mu^2)$$

$$\frac{\partial^2 \ln c}{\partial w_2^2} = -\frac{n}{2w_2^2} - \frac{nw_1^2}{2w_2^3} = -n(2\sigma^4 + 4\mu^2\sigma^2)$$

故　$\mathrm{Var}[T_2(X)] = n(2\sigma^4 + 4\mu^2\sigma^2)$

$$\frac{\partial^2 \ln c}{\partial w_1 \partial w_2} = -\frac{nw_1}{2w_2^2} = -2n\mu\sigma^2,\text{故 } \mathrm{Cov}[T_1(X), T_2(X)] = 2n\mu\sigma^2.$$

定理 1.15　设 $X=(X_1,\cdots,X_n)$是来自指数型分布的一个样本,其联合密度函数为

$$p_w(x) = c(\boldsymbol{w})\exp\left\{\sum_{j=1}^{k} w_j T_j(x)\right\} h(x)\boldsymbol{w} \in \Omega$$

其中$(T_1(x),\cdots,T_k(x))$如(1.74)所示,Ω 为 $\mathbf{R}^k$ 的子集,假如 Ω 有内点,则统计量$(T_1(x),\cdots,T_k(x))$是完备统计量.

由于证明较为冗长,这里就不再叙述了,有兴趣的读者可参阅陈希孺的专著[1].用这个定理可以验证许多常见统计量的完备性.

参 考 文 献

1　陈希孺. 数理统计引论,北京:科学出版社,1981

2　成平,陈希孺,陈桂景,吴传义. 参数估计. 上海:上海科技出版社,1985

3　张尧庭,方开泰. 多元统计分析引论. 北京:科学出版社,1982

4　严士健,刘秀芳. 测度与概率. 北京:北京师范大学出版社,1994

5　茆诗松,王静龙. 数理统计. 上海:华东师范大学出版社,1990

6　Bickel P J, Doksum K A. Mathematical Statistics. San Fancisco: Holden-Day Inc., 1977. 中译本:李泽慧,王嘉澜,林亨泽. 数理统计. 兰州:兰州大学出版社,1991

7　Singer S P K. Large sample methods in statistics. New York: Chapman-Hall, 1993

8　Halmos P R, Savage L J. Application of the Radom-Nikodym

theorem to sufficient statistics. Ann Math Statistics, 1949, 20. 225～241

9 Cramer H. Mathematical methods of Statistics. Princeton Univ. Press, 1946. 中译本：魏宗舒，郑扑，吴锦译. 统计学数学方法，上海：上海科学技术出版社，1966

10 陈善林，张浙. 统计发展史. 北京：立信会计图书用品社，1987

习 题 一

1.1 写出下列统计问题的统计结构.

(1)在测量未知量 μ 的问题中，若已知测量仪器有 $+0.1$ 的偏差，而误差仍是方差已知的正态变量；

(2)在(1)中，若已知测量仪器有正的偏差，但偏差大小未知，其它不变.

1.2 写出下列统计问题的统计结构.

(1)一地质师在一老河床测量 n 个卵石的最大直径，若已知直径的对数服从均值为 μ 和方差为 σ^2 的正态分布；

(2)一昆虫产出的卵数服从均值为 λ 的 Poisson 分布，卵一旦产出，每个卵能孵出幼虫的概率为 θ，并且每个卵的孵化是相互独立的，一位昆虫学家对 n 个昆虫观察所产生的卵数 X 和孵化出的幼虫数 Y；

(3)已知人的体重 Y 与身高 x 有关，一般认为较高的人体重较重，且有线性趋势，但同样身高的人的体重且不会都相同. 如今测量 n 个人的体重 y_i 和身高 x_i，并设有如下关系

$$y_i = a + bx_i + \varepsilon_i, \quad i = 1, \cdots, n$$

其中诸 ε_i 相互独立，$E(\varepsilon_i)=0$，$Var(\varepsilon_i)=\sigma^2$，$i=1,\cdots,n$. 而 a，b 和 σ^2 都是未知量.

1.3 设$(\mathscr{X},\mathscr{B},\mathscr{P})$是一个统计结构，证明：

(1)若 $\mathscr{X}$ 只含有不多于可数个元素，则此结构是可控的；

(2)若 $\mathscr{P}$ 只含有不多于可数个分布，则此结构是可控的.

1.4 设随机变量 $X \sim Ga(\alpha,\lambda)$，若 c 为正实数，证明 $cX \sim$

$Ga(\alpha,\lambda/c)$，若 $X \sim Ga(5,0.01)$，计算概率 $P(X>200)$.

1.5　设随机变量 $X \sim Ga(\alpha,\lambda)$，则 $Y=X+\mu$ 服从三参数 Gamma 分布，其中 μ 称为门限参数，请写出 Y 的密度函数，并计算 $E(Y)$ 和 $Var(Y)$.

1.6　设随机变量 $X \sim LN(\mu,\sigma^2)$，则 $Y=X+\alpha$ 服从三参数对数正态分布，其中 α 称为门限参数，请写出 Y 的密度函数，并计算 $E(Y)$ 和 $Var(Y)$.

1.7　设随机变量 $X \sim Exp(\lambda)$，证明：对任意非负实数 t 与 s，有（无记忆性）

$$P(X > t + s | X > s) = P(X > t)$$

1.8　设随机变量 $X \sim \chi^2(2)$，证明：自由度为 2 的 χ^2 分布的 α 分位数（$0<\alpha<1$）为 $-2\ln(1-\alpha)$.

1.9　χ^2 分布函数表和 χ^2 分布的分位数表往往只对自由度 $n \leqslant 30$ 给出，当 $n>30$ 时可用正态分布近似，计算

(1)自由度为 35 的 χ^2 变量大于 45 的概率；

(2)自由度为 40 的 χ^2 分布的 0.05 分位数.

1.10　设随机变量 $X \sim Ga(\alpha,\lambda)$，则 $Y=X^{-1}$ 服从倒 Gamma 分布，请写出 Y 的密度函数，并计算 $E(Y)$ 和 $Var(Y)$.

1.11　若随机变量 $X+c$ 和 $-X+c$ 有相同的分布，则称随机变量的分布关于点 c 对称. 如果 X 有密度函数 $p(x)$，则其关于点 c 对称的充要条件是

$$p(x-c) = p(c-x), \qquad \forall\, x \in \mathbf{R}$$

1.12　设随机变量 $X \sim Be(a,b)$，证明：$Y=1-X \sim Be(b,a)$.

1.13　设随机变量 $X \sim Z(a,b)$，证明：$Y=X^{-1} \sim Z(b,a)$.

1.14　设 $F_\alpha(n_1,n_2)$ 是自由度为 n_1,n_2 的 F 分布的 α 分位数，证明：$F_{1-\alpha}(n_1,n_2)=[F_\alpha(n_2,n_1)]^{-1}$.

1.15　证明下述结论.

(1)设 $F(x)$ 为连续随机变量 X 的分布函数，则

$$Y = F(X) \sim U(0,1)$$

(2)设 $Y \sim U(0,1)$，则 $u=-2\ln Y \sim \chi^2(2)$；

(3)设 $X_1,\cdots,X_n$ 是连续随机变量 X 的 n 次观察值,$F(x)$是 X 的分布函数,则 $-2\sum_{i=1}^{n}\ln F(X_i)\sim\chi^2(2n)$.

1.16 验证:自由度为 2 和 $2k$ 的 F 分布的 α 分位数是

$$F_\alpha(2,2k)=k[(1-\alpha)^{-1/k}-1].$$

1.17 设 $X_1,\cdots,X_{n_1}$是来自 $N(\mu,\sigma^2)$的一个样本,$Y_1,\cdots,Y_{n_2}$是来自 $N(0,\sigma^2)$的一个样本,且诸 X_i 与诸 Y_j 相互独立,证明:

$$F=\frac{\left(X_1^2+\cdots+X_{n_1}^2\right)/n_1}{\left(Y_1^2+\cdots+Y_{n_2}^2\right)/n_2}\sim F(n_1,n_2,\gamma)$$

其中 $F(n_1,n_2,\gamma)$称为非中心参数为 $\gamma=n_1\mu^2/(2\sigma^2)$的非中心 F 分布,其密度函数在 $u\leqslant 0$ 为零,在 $u>0$ 为

$$p(u;n_1,n_2,\gamma)=\frac{n_1^{\frac{n_1}{2}}n_2^{\frac{n_2}{2}}\mathrm{e}^{-\gamma}u^{\frac{n_1}{2}-1}}{\Gamma\left(\frac{n_1}{2}\right)}\cdot$$

$$\sum_{m=0}^{\infty}\frac{\Gamma\left(\frac{n_1+n_2}{2}+m\right)}{\Gamma\left(\frac{n_2}{2}+m\right)}\cdot\frac{(\gamma n_1 u)^m}{m!(n_2+n_1u)^{\frac{n_1+n_2}{2}+m}}$$

并求其期望和方差.

1.18 定义在$(\mathbf{R},\mathscr{B}_{\mathbf{R}})$上用密度函数

$$p(x;\mu,\sigma^2,\nu)=\frac{\Gamma\left(\frac{\nu+2}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}\sigma}\left[1+\frac{1}{\gamma}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\frac{\gamma+1}{2}}$$

表示的概率分布称为一般 t 分布,其中 $\mu\in\mathbf{R}$ 称为位置参数,$\sigma>0$ 称为尺度参数,$\nu>0$ 称为自由度.证明:

$$\mathrm{E}(X)=\mu,\quad \nu>1$$

$$\mathrm{Var}(X)=\frac{\nu\sigma^2}{\nu-2},\quad \nu>2$$

1.19 设$(X_1,\cdots,X_r)\sim M(n,p_1,\cdots,p_r)$,$r\geqslant 3$,求 $Y=X_1+X_2$ 的分布.

1.20 设$(X_1,\cdots,X_r)\sim M(n,p_1,\cdots,p_r)$,在给定 $X_2=n_2$ 的条件

下，求 X_1 的条件分布．

1.21　设 $\boldsymbol{X}\sim N_n(\boldsymbol{\mu},\boldsymbol{\Sigma})$，$\boldsymbol{C}$ 为任一 $r\times n$ 阶阵，证明

$$\boldsymbol{CX}\sim N_r(\boldsymbol{C\mu},\boldsymbol{C\Sigma C}')$$

1.22　设

$$X=\begin{pmatrix}X_1\\X_2\end{pmatrix}\sim N_n\left(\begin{pmatrix}\mu_1\\\mu_2\end{pmatrix},\begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\\boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}\right)$$

其中 $\boldsymbol{X}_1$ 与 μ_1 都是 k 维向量，$\boldsymbol{\Sigma}_{11}$为 k 阶方阵，$\boldsymbol{\Sigma}_{12}$，$\boldsymbol{\Sigma}_{21}$，$\boldsymbol{\Sigma}_{22}$为相应矩阵，且 $|\boldsymbol{\Sigma}_{22}|\neq 0$，证明

(1)$\boldsymbol{X}_1\sim N_k(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_{11})$；

(2)$\boldsymbol{X}_1$ 与 $\boldsymbol{X}_2$ 相互独立的充要条件是 $\boldsymbol{\Sigma}_{12}=0$；

(3)在给定 $\boldsymbol{X}_2=\boldsymbol{x}_2$ 下，$\boldsymbol{X}_1$ 的条件分布是

$$N_k(\mu_1+\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2-\mu_2),\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}'_{12})$$

1.23　设 $X_1\sim N(0,1)$，$X_2\sim N(0,4)$，且 X_1 与 X_2 独立，求 $Y_1=X_1+X_2$ 和 $Y_2=X_1-X_2$ 的联合分布．

1.24　证明：一个 n 元随机变量 $\boldsymbol{X}=(X_1,\cdots,X_n)$服从 n 元正态分布的充要条件是 $\boldsymbol{X}$ 的诸分量的任一个线性组合都服从一元正态分布．

1.25　设$\begin{pmatrix}X\\Y\end{pmatrix}$为二元正态变量，其分布为 $N_2\left(\begin{pmatrix}0\\1\end{pmatrix},\begin{pmatrix}4 & 1\\1 & 1\end{pmatrix}\right)$写出给定 X 下，Y 的条件分布和给定 Y 下，X 的条件分布．

1.26　设$\begin{pmatrix}X_1\\X_2\end{pmatrix}\sim N_2\left(\begin{pmatrix}\mu_1\\\mu_2\end{pmatrix},\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\\rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right)$，证明

$$Y_1=\frac{X_1-\mu_1}{\sigma_1},Y_2=\frac{1}{\sqrt{1-\rho^2}}\left(\frac{X_2-\mu_2}{\sigma_2}-\rho\frac{X_1-\mu_1}{\sigma_1}\right)$$

为相互独立的标准正态变量．

1.27　设 $X_1\sim Ga(\alpha_1,\lambda)$，$X_2\sim Ga(\alpha_2,\lambda)$，且 X_1 与 X_2 独立，证明：

(1)$Y_1=X_1+X_2$ 与 $Y_2=X_1/(X_1+X_2)$独立，且 $Y_2\sim Be(\alpha_1,\alpha_2)$；

(2)$Y_1=X_1+X_2$ 与 $Y_3=X_1/X_2$ 独立，且 $Y_3\sim Z(\alpha_1,\alpha_2)$．

1.28　设 $X_1\sim N(0,\sigma_1^2)$，$X_2\sim N(0,\sigma_2^2)$，且 X_1 与 X_2 独立，寻找 $Y_1=X_1+X_2+a$ 与 $Y_2=X_1-X_2+b$ 的联合分布．

1.29　设 X_1 与 X_2 是相互独立且服从同一指数分布 $Exp(\lambda)$的随

机变量，寻找 $Y_1=X_1-X_2$ 和 $Y_2=X_2$ 的联合密度函数和 Y_1 的边际密度函数.

1.30　设总体 X 有密度函数

$$p(x)=\begin{cases}2x, & 0<x<1\\ 0, & \text{其它}\end{cases}$$

从该总体随机抽取一个容量为 4 的样本，计算概率 $P(X_{(3)}>0.5)$.

1.31　设 $X_1,\cdots,X_n$ 是来自 Weibull 分布的一个样本. 其分布函数为：当 $x\leqslant 0$ 时 $F(x)=0$，当 $x>0$ 时

$$F(x)=1-\exp\left\{-\left(\frac{x}{\eta}\right)^m\right\}$$

其中 $m>0$ 为形状参数，$\eta>0$ 为尺度参数，证明：$Y=\min(X_1,\cdots,X_n)$ 仍服从 Weibull 分布.

1.32　设 $X_{(1)}$ 和 $X_{(n)}$ 分别为容量 n 的样本的最小和最大次序统计量，证明：极差 $R_n=X_{(n)}-X_{(1)}$ 的分布函数

$$F_{R_n}(x)=n\int_{-\infty}^{\infty}[F(y+x)-F(y)]^{n-1}p(y)\mathrm{d}y$$

其中 $F(y)$ 与 $p(y)$ 分布为总体的分布函数与密度函数.

1.33　当总体为指数分布 $Exp(\lambda)$ 和均匀分布 $U(0,1)$ 时，分别寻求容量为 n 的样本极差 R_n 的分布.

1.34　设总体 X 的分布函数和密度函数分别为 $F(x)$ 和 $p(x)$，又设 $X_{(1)}$ 和 $X_{(n)}$ 分别为从该总体抽取容量为 n 的样本的极小和极大次序统计量. 寻求样本极差 $W=X_{(n)}-X_{(1)}$ 和样本中程 $M=(X_{(1)}+X_{(n)})/2$ 的联合分布及各自的边际分布.

1.35　设 $X_{(1)},\cdots,X_{(n)}$ 是某样本的次序统计量，当剔去其中前 k 个和后 k 个变量后，用剩下 $n-2k$ 个次序统计量计算平均

$$T_{n,k}=\frac{1}{n-2k}\sum_{i=k+1}^{n-k}X_{(i)}$$

这个平均称为切断平均，其中 $k<n/2$. 请指出切断平均的期望与方差存在的条件.

1.36　设 $X_1,\cdots,X_n$ 是来自总体分布函数 $F(x)$ 的一个样本，$F_n(x)$ 为其经验分布函数

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \leqslant x\}}$$

证明：

$$\sqrt{n}\,[F_n(x) - F(x)] \xrightarrow{L} N(0, F(x)[1-F(x)])$$

1.37　设 $X_1,\cdots,X_{n_1},X_{n_1+1},\cdots,X_{n_1+n_2}$ 是来自标准正态分布总体 $N(0,1)$ 的一个样本，大家知道

$$F = \frac{\sum_{i=1}^{n_1} X_i^2/n_1}{\sum_{i=n_1+1}^{n_1+n_2} X_i^2/n_2} \sim F(n_1,n_2)$$

证明：当 n_1 较小和 n_2 较大时，可用 $\chi^2(n_1)$ 来近似 $F(n_1,n_2)$.

1.38　设 $X_1,\cdots,X_n$ 是来自二点分布 $b(1,\theta)$ 的一个样本，其中 $0<\theta<1$，现要求样本均值 $\overline{X}$ 的函数

$$g_1(\overline{X}) = (\overline{X})^{-1}$$

$$g_2(\overline{X}) = \sqrt{\overline{X}(1-\overline{X})}$$

的渐近分布.

1.39　寻求 t 分布的 Edgeworth 展开式.

1.40　在下列密度函数下分别寻求容量为 n 的样本中位数 $m_{0.5}$ 的渐近分布.

(1) $p(x)=\begin{cases} 2x, & 0<x<1, \\ 0, & \text{其它}; \end{cases}$

(2) $p(x)=\dfrac{1}{\pi(1+x^2)}$;

(3) $p(x)=\dfrac{\lambda}{2}\mathrm{e}^{-\lambda|x|}$;

(4) $p(x)=\dfrac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\dfrac{(x-\mu)^2}{2\sigma^2}\right\}$.

1.41　设 $X_1,\cdots,X_n$ 是来自二点分布 $b(1,\theta)$ 的一个样本，证明

$$T_k = (X_1 + \cdots + X_k, X_{k+1}, \cdots, X_n),\quad X = 1,2,\cdots,n$$

都是 θ 的充分统计量. 若对 T_k 给定 $(t, t_{k+1}, \cdots, t_n)$，请设计一个随机试

验，它能产生与原样本同分布的新样本.

1.42 设 $X_1,\cdots,X_n$ 是来自二项分布 $b(m,\theta)$ 的一个样本，证明：$T_n=\sum_{i=1}^n X_i$ 是 θ 的充分统计量.

1.43 设有一个独立的贝努里试验序列，其每次成功概率均为 θ，又设 X 是该序列中首次成功发生前的失败次数则其分布为

$$P_\theta(X=x)=(1-\theta)^x\theta,\qquad x=0,1,\cdots$$

这个分布称为几何分布. 若 $X_1,\cdots,X_n$ 是来自该几何分布的一个样本.

(1) 寻求统计量 $T_n=\sum_{i=1}^n X_i$ 的分布；

(2) 证明：T_n 是充分统计量.

1.44 设 $X_1,\cdots,X_n$ 是来自如下密度函数的一个样本，分别求未知参数 θ 的充分统计量.

(1)(幂分布)$p_\theta(x)=\theta x^{\theta-1}$, $0<x<1$, $\theta>0$;

(2)(Pareto 分布)$p_\theta(x)=\theta a^\theta/x^{(\theta+1)}$, $x>a,\theta>0(a>0$ 已知)；

(3)(拉普拉斯分布)$p_\theta(x)=\frac{1}{\theta}\mathrm{e}^{-|x|/\theta}$, $-\infty<x<\infty$, $\theta>0$.

1.45 设 $X_1,\cdots,X_n$ 是来自均匀分布 $U(\theta_1,\theta_2)$ 的一个样本，其 $-\infty<\theta_1<\theta_2<\infty$. 证明：$(X_{(1)},X_{(n)})$ 是参数 (θ_1,θ_2) 的充分统计量，其中 $X_{(1)}$ 与 $X_{(n)}$ 分别是该样本的最小与最大次序统计量.

1.46 设 $X_1,\cdots,X_n$ 是来自 Gamma 分布族 $\{Ga(\alpha,\lambda):\alpha>0,\lambda>0\}$ 的一个样本，寻求 (α,λ) 的充分统计量.

1.47 设 $X_1,\cdots,X_n$ 是来自 Beta 分布族 $\{Be(a,b),a>0,b>0\}$ 的一个样本，寻求 (a,b) 的充分统计量.

1.48 设 $X_1,\cdots,X_r$ 服从参数为 $p_1,\cdots,p_r$ 的多项分布 $M(n;p_1,\cdots,p_r)$，设 $\left(X_{1_j},\cdots,X_{r_j}\right)$, $j=1,\cdots,m$ 是来自多项分布的一个多维样本，证明：$\left(\sum_{j=1}^m X_{1_j},\cdots,\sum_{j=1}^m X_{r_j}\right)$ 是 $(p_1,\cdots,p_r)$ 的充分统计量.

1.49 设二维随机变量 $\boldsymbol{X}=\begin{pmatrix}X_1\\X_2\end{pmatrix}$ 服从二元正态分布，其均值向量为零向量，协方差阵为

$$\begin{pmatrix}\sigma^2+r^2 & \sigma^2-r^2\\ \sigma^2-r^2 & \sigma^2+r^2\end{pmatrix},\quad \sigma>0,r>0$$

证明：二维统计量 $T=((X_1+X_2)^2,(X_1-X_2)^2)$ 是该二元正态分布族的充分统计量.

1.50　设 $Y_1,\cdots,Y_n$ 是相互独立的随机变量，且

$$Y_i\sim N(\alpha+\beta x_i,\sigma^2),\quad i=1,\cdots,n$$

其中 $-\infty<\alpha,\beta<\infty,\sigma>0$ 是未知参数，而 $X_1,\cdots,X_n$ 是已知量.

证明：$\left(\sum_{i=1}^n X_i,\sum_{i=1}^n X_iY_i,\sum_{i=1}^n Y_i^2\right)$ 是正态分布族

$$\{N(\alpha+\beta x,\sigma^2);\ -\infty<\alpha,\beta<\infty,\sigma>0\}$$

的充分统计量.

1.51　设 $\begin{pmatrix}X_i\\Y_i\end{pmatrix}$，　$i=1,\cdots,n$ 是来自正态分布族

$$\left\{N\left(\begin{pmatrix}\theta_1\\\theta_2\end{pmatrix},\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\\rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right),-\infty<\theta_1,\theta_2<\infty,\sigma_1,\sigma_2>0,|\rho|\leqslant 1\right\}$$

的一个二维样本，寻求该分布族的充分统计量.

1.52　检查下列分布族的完备性

(1) Poisson 分布族；

(2) 几何分布族 $\{p_\theta(x)=\theta(1-\theta)^x,x=0,1,\cdots,0<\theta<1\}$；

(3) 均匀分布族 $\{U(0,\theta),\theta>0\}$；

(4) 均匀分布族 $\{U(\theta,\theta+1,\theta\in\mathbf{R})\}$；

(5) 正态分布族 $\{N(\mu,\sigma^2),\mu\in\mathbf{R},\sigma\in\mathbf{R}^+\}$；

(6) Gamma 分布族 $\{Ga(\alpha,\lambda),\alpha\in\mathbf{R}^+,\lambda\in\mathbf{R}^+\}$；

(7) Cauchy 分布族 $\left\{p_\theta(x)=\dfrac{\theta}{\pi(\theta^2+x^2)},\theta>0\right\}$.

1.53　设 $X_1,\cdots,X_n$ 是来自均匀分布 $U(0,\theta)$ 的一个样本，其中 $\theta\in\mathbf{R}^+$，证明：$X_{(n)}=\max(X_1,\cdots,X_n)$ 是完备统计量.

1.54　考察如下的幂级数分布

$$P_\theta(X=x)=\frac{a_x}{f(\theta)}\theta^x,\quad x=c,c+1,\cdots$$

设 $X_1,\cdots,X_n$ 是来自此分布的一个样本，又设 $T=X_1+\cdots+X_n$，

证明：(1) T 的分布具有同样形式；(2) T 是完备统计量.

1.55　下列分布族中哪一个是指数族.

(1) $p_\theta(x)=2x/\theta^2,\quad 0<x<\theta$;

(2) $p_\theta(x)=1/9,\quad x=\theta+0.1,\quad \theta+0.2,\cdots,\theta+0.9$;

(3) Poisson 分布族;

(4) Gamma 分布族;

(5) 正态分布族$\{N(\theta,\theta^2),\quad \theta>0\}$;

(6) $p_\theta(x)=2(x+\theta)/(1+2\theta),\quad 0<x<1,\theta>0$;

(7) Beta 分布族;

(8) $p_\theta(x)=\theta(1-\theta)^x,\quad x=0,1,\cdots,0<\theta<1$;

(9) 均匀分布族$U(\theta-0.5,\theta+0.5),\quad \theta\in\mathbf{R}$;

(10) 对数正态分布族;

(11) 多项分布族;

(12) $p_{\mu,\sigma}(x)=\frac{1}{\sigma}\exp\left\{-\frac{x-\mu}{\sigma}\right\},\quad \mu\in\mathbf{R},\sigma>0,x>\mu$.

1.56 把下列分布的密度函数写成指数族的标准形式,并指出其自然参数空间.

(1) 二项分布;

(2) Poisson 分布;

(3) Gamma 分布;

(4) 二元正态分布.

第二章　点　估　计

上一章，我们介绍了统计结构的概念，并指出，与概率论相比，统计的重要特点在于分布的不确定性．实际中，在得到样本以后，我们最关心的问题之一便是想知道是分布族中的哪一个分布产生出此样本，即要从样本推断总体分布或其各种特征数．英国著名统计学家 R. A. Fisher 把统计推断归纳为三个方面：抽样分布，参数估计与假设检验，其中参数估计又分为点估计与区间估计．抽样分布已在第一章讨论过，以后几章还会经常涉及，本章主要讨论点估计，后面两章将分别讨论假设检验和区间估计．

§2.1　估计与优良性

§2.1.1　参数及其估计

在讲述参数估计之前，我们首先必须清楚什么是参数．如果统计结构的分布族由 $\mathscr{P}=\{P_\theta,\theta\in\Theta\}$ 给出，其均值，方差等特征数都是 θ 的实值函数，一般地，任何定义在 Θ 上的实值函数都可以称为参数，但参数的概念并不局限于参数统计结构，在非参数统计结构中也有参数．譬如，若 X 的分布族为 $\mathscr{P}=\{P\}$，则均值 $\mathrm{E}_P X$，方差 $\mathrm{Var}_P(X)$ 等分布的特征数也都是参数．由此，我们可以给出如下定义．

定义 2.1　定义在 $\mathscr{P}$ 上的一个实值泛函 $g(P)$ 称为参数，而 $(\mathscr{X},\mathscr{B},\mathscr{P})$ 上的用来估计 $g(P)$ 的实值统计量称为 $g(P)$ 的点估计量，简称估计．

简单起见，常以 θ 表示参数，以 $\hat\theta=\hat\theta(X)$ 表示其点估计．由定义可以看出，估计的概念相当广泛，任何人都可以给出估计．如果不对估计

的好坏加以明确.估计是没有意义的.下面,我们讨论估计的好坏标准.

§2.1.2 均方误差

假设我们用 $\hat{\theta}(X)$ 估计 θ,评价该估计好坏的一个自然度量是 $|\hat{\theta}(X)-\theta|$,由于 θ 是未知的,样本 X 又具有随机性,直接使用这种自然度量在实际中是不可行的,为排除样本随机性的影响,可以对它求期望,出于数学处理上的方便考虑,最常用的标准是由下式给出的均方误差.

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathrm{E}(\hat{\theta}(X)-\theta)^2 \tag{2.1}$$

自然,我们希望估计的均方误差越小越好.如果能找到 θ 的一个估计 $\hat{\theta}^*$,使得对所有估计 $\hat{\theta}$ 有

$$\mathrm{MSE}_\theta(\hat{\theta}^*) \leqslant \mathrm{MSE}_\theta(\hat{\theta}), \quad \forall\, \theta \in \Theta \tag{2.2}$$

则 $\hat{\theta}^*$ 应是最好的估计.遗憾的是,这样的 $\hat{\theta}^*$ 是不存在的,因为倘若这样的 $\hat{\theta}^*$ 存在,对任一 $\theta_0 \in \Theta$,取 $\hat{\theta}_0(X) \equiv \theta_0$,则 $\mathrm{MSE}_{\theta_0}(\hat{\theta}_0)=0$,从而由(2.2)有 $\mathrm{MSE}_{\theta_0}(\hat{\theta}^*)=\mathrm{E}_{\theta_0}(\hat{\theta}^*(X)-\theta_0)^2=0$,这表明

$$\hat{\theta}^*(X) = \theta_0, \quad \text{a.s.}\ P_{\theta_0}$$

由 θ_0 的任意性,故这样的 $\hat{\theta}^*$ 找不到.

由此可见,使均方误差一致达到最小的最优估计是不存在的.那末如何寻找好的估计呢?办法之一是先对估计提出一些合理性要求,把那些诸如 $\hat{\theta}(X) \equiv \theta_0$ 的不合理估计排除在外,然后在满足这种合理性要求的估计类中寻找好的估计.无偏性便是一种最常用的合理性要求.

§2.1.3 无偏性

若用 $\hat{\theta}=\hat{\theta}(X)$ 估计 θ,则由简单的数学推导易知

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + (\mathrm{E}(\hat{\theta})-\theta)^2 \tag{2.3}$$

其中 $\mathrm{E}(\hat{\theta})-\theta$ 称为估计 $\hat{\theta}$ 的偏差.如果偏差等于0,就是所谓的无偏估计.

定义 2.2 设 $(\mathscr{X},\mathscr{B},\{P_\theta,\theta\in\Theta\})$ 为可控参数统计结构,$g(\theta)$ 是未知参数,$X=(X_1,\cdots,X_n)$ 是来自该统计结构的一个样本,若用 $\hat{g}(X)$

估计 $g(\theta)$，且

$$E_\theta(\hat{g}(X)) = g(\theta), \quad \forall\, \theta \in \Theta$$

则称 $\hat{g}(X)$ 为 $g(\theta)$ 的无偏估计.

无偏性体现了一种频率思想. 只有在大量重复使用时，无偏性才有意义. 譬如，要估计某批产品的合格率 θ，从中抽取 n 个产品进行检查，发现有 X 个合格，容易验证，在假定 $X \sim b(n,\theta)$ 下，X/n 是 θ 的无偏估计. 然而对一次具体观测值 x 来说，x/n 要么等于 θ 要么不等，此时无偏性显得没有意义；然而，如果问题改为某一工厂每天都对其生产的产品进行抽检，若假定生产过程相对稳定，则估计的无偏性要求便是合理的. 比如用 X/n 估计 θ，对每一天而言，该估计可能偏大也可能偏小，但在一段较长的时期内，x/n 平均来说在 θ 的周围波动.

例 2.1　设 $X_1,\cdots,X_n$ 是来自 $N(\mu,\sigma^2)$ 的一个样本，μ 和 σ^2 是未知参数，其常用估计分别是样本均值

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.4}$$

和样本方差

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{2.5}$$

易知

$$E(\overline{X}) = \mu$$

又由于 $ns_n^2/\sigma^2 \sim \chi_{n-1}^2$，故

$$E(s_n^2) = \frac{\sigma^2}{n}E\left(\frac{ns_n^2}{\sigma^2}\right) = \frac{n-1}{n}\sigma^2$$

从而 $\overline{X}$ 是 μ 的无偏估计，而 s_n^2 则不是 σ^2 的无偏估计. 但对 s_n^2 略加修正就可得到 σ^2 的一个无偏估计

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{2.6}$$

s_n^2 和 s^2 都是 σ^2 的常用估计，在实际中，s^2 用得更多，这是因为当样本容量不是很大时，s_n^2 常过小估计 σ^2. 如果样本容量较大，s_n^2 与 s^2 很接近，此时 $E(s_n^2) - \sigma^2$ 接近 0，s_n^2 称为 σ^2 的渐近无偏估计.

定义 2.3　设 $\hat{g}_n = \hat{g}_n(X_1,\cdots,X_n)$ 是 $g(\theta)$ 的估计量，若

$$\lim_{n\to\infty} \mathrm{E}_\theta(\hat{g}_n) = g(\theta), \quad \forall\, \theta \in \Theta$$

则称 $\hat{g}_n$ 为 $g(\theta)$的渐近无偏估计.

§2.1.4 相合性

估计量是与样本容量有关的,假设用 $\hat{\theta}_n=\hat{\theta}_n(X_1,\cdots,X_n)$估计 θ,我们不可能做到对某一 n,MSE($\hat{\theta}_n$)对所有 $\theta\in\Theta$ 任意小,但当 $n\longrightarrow\infty$时通常可以做到这一点,这便是相合性的概念.

定义 2.4 设 $\hat{\theta}_n=\hat{\theta}_n(X_1,\cdots,X_n)$是 θ 的估计.如果当 $n\longrightarrow\infty$时,有

$$\hat{\theta}_n \xrightarrow{P} \theta,$$

则称 $\hat{\theta}_n$ 是 θ 的(弱)相合估计.进一步,如果

$$\hat{\theta}_n \longrightarrow \theta, \text{a. s.}$$

则称 $\hat{\theta}_n$ 为 θ 的强相合估计.

显然,强相合性可以推出弱相合性.在统计研究中,弱相合性便已足够,在以后的讨论中,我们所指的相合性均是指弱相合性.

相合性被认为是对估计的一个最基本要求.如果一个估计量,无论做多少次试验或有多少个观测值,它都不能把要估计的参数估计到任意指定的精度.那末这个估计是很值得怀疑的.通常,不满足相合性要求的估计一般不予考虑.

例 2.2 设 $X_1,\cdots,X_n$ 是来自 $U(0,\theta)$的一个样本,最大次序统计量 $X_{(n)}$是 θ 的常用估计.由§1.3 节中(1.39)式知 $X_{(n)}$的密度函数为

$$p(t;\theta) = nt^{n-1}\theta^{-n}, \quad 0 < t < \theta$$

易求出 $\mathrm{E}(X_{(n)})=n\theta/(n+1)$,因此 $X_{(n)}$不是 θ 的无偏估计,但它是 θ 的渐近无偏估计.另外,由于对任意的 $\varepsilon>0$,

$$\begin{aligned} P_\theta(|X_{(n)} - \theta| \geqslant \varepsilon) &= P_\theta(X_{(n)} \leqslant \theta - \varepsilon) \\ &= \int_0^{\theta-\varepsilon} \frac{nt^{n-1}}{\theta^n}\mathrm{d}t \\ &= \left(\frac{\theta-\varepsilon}{\theta}\right)^n \longrightarrow 0 \ (n\to\infty) \end{aligned}$$

因此,$X_{(n)}$是 θ 的相合估计.

证明估计的相合性一般可应用大数定律或直接由定义来证. 下述定理在判断相合估计时是很有用的.

定理 2.1　设 $T_{jn}=T_{jn}(X_1,\cdots,X_n)$ 是 $g_j(\theta)$ 的相合估计, $j=1,\cdots,k$. 函数 $h(\cdot)$ 在 $(g_1(\theta),\cdots,g_k(\theta))$ 处连续, 则 $h(T_{1n},\cdots,T_{kn})$ 是 $h(g_1(\theta),\cdots,g_k(\theta))$ 的相合估计.

证明: 因 $h(\cdot)$ 在 $(g_1(\theta),\cdots,g_k(\theta))$ 处连续, 故 $\forall\ \varepsilon>0$, $\exists\ \eta>0$, 当 $|g_j(\theta)-y_j|<\eta$, $j=1,\cdots,k$ 时, 有

$$|h(y_1,\cdots,y_k)-h(g_1(\theta),\cdots,g_k(\theta))|<\varepsilon$$

又 T_{jn} 是 $g_j(\theta)$ 的相合估计, 对此 $\varepsilon>0$, $\exists\ N_j$, 当 $n>N_j$ 时

$$Pr(|T_{jn}-g_j(\theta)|\geqslant\eta)<\varepsilon/k$$

取 $N=\max(N_1,\cdots,N_k)$, 当 $n>N$ 时,

$$\begin{aligned}
&Pr(|T_{jn}-g_j(\theta)|<\eta, j=1,\cdots,k)\\
&=1-Pr(\text{至少有一个 } i, |T_{in}-g_i(\theta)|\geqslant\eta)\\
&\geqslant 1-\sum_{j=1}^{k}Pr(|T_{jn}-g_j(\theta)|\geqslant\eta)>1-\varepsilon
\end{aligned}$$

所以

$$\begin{aligned}
&Pr(|h(T_{1n},\cdots,T_{kn})-h(g_1(\theta),\cdots,g_k(\theta))|<\varepsilon)\\
&\geqslant Pr(|T_{jn}-g_j(\theta)|<\eta, j=1,\cdots,k)\\
&>1-\varepsilon
\end{aligned}$$

证毕.

例 2.3　设 $X_1,\cdots,X_n$ 是来自密度函数为

$$p(x;\theta)=\theta(\theta+1)x^{\theta-1}(1-x),\quad 0<x<1,\theta>0$$

的一个样本. 因

$$\mathrm{E}(X_1)=\int_0^1 xp(x;\theta)\mathrm{d}x=\theta/(\theta+2)\triangleq\eta$$

由强大数定律有

$$\overline{X}\longrightarrow\eta,\quad \text{a.s.}$$

而 $\theta=\dfrac{2\eta}{1-\eta}\triangleq h(\eta)$ 是 η 的连续函数, 由定理 2.1 知, $h(\overline{X})=\dfrac{2\overline{X}}{1-\overline{X}}$ 是 θ 的相合估计.

必须指出，相合性只是反映了当 $n \longrightarrow \infty$ 时估计量的性质，而对任意有限的 n，相合性是没有意义的. 相合性本身不能说明为使 $\hat{\theta}_n$ 达到一定精度 n 必须至少为多少，事实上，相合估计可以不止一个，它们之间是有差异的，这种差异往往可由估计量的渐近分布的渐近方差反映出来. 最常用的渐近分布是正态分布.

§2.1.5 渐近正态性

定义 2.5 设 $\hat{\theta}_n = \hat{\theta}_n(X_1, \cdots, X_n)$ 是 θ 的估计，如果存在 $\sigma_n^2(\theta)$，满足

$$(\hat{\theta}_n - \theta)/\sigma_n(\theta) \xrightarrow{L} N(0,1) \tag{2.7}$$

则称 $\hat{\theta}_n$ 是 θ 的渐近正态估计，$\sigma_n^2(\theta)$ 称为 $\hat{\theta}_n$ 的渐近方差. 记 $\hat{\theta}_n \sim AN(\theta, \sigma_n^2(\theta))$.

显然，如果满足(2.7)式的 $\sigma_n^2(\theta)$ 存在，则它是不唯一的. 由 Slutskey 定理(定理 1.5)，若 $\tilde{\sigma}_n^2(\theta)$ 满足

$$\tilde{\sigma}_n(\theta)/\sigma_n(\theta) \longrightarrow 1$$

则有 $(\hat{\theta}_n - \theta)/\tilde{\sigma}_n(\theta) \xrightarrow{L} N(0,1)$. 一般 $\sigma_n^2(\theta)$ 可取为 $\hat{\theta}_n$ 的方差.

由中心极限定理，若 $X_1, \cdots, X_n$ 是独立同分布随机变量，$\mathrm{E}(X_1) = \theta$，$\mathrm{Var}(X_1) = \sigma^2 < \infty$，则 $\sqrt{n}(\overline{X} - \theta)/\sigma \sim AN(0,1)$. 由此可见，渐近正态估计的渐近方差往往具有 $\frac{1}{n}$ 的阶，即

$$n\sigma_n^2(\theta) \longrightarrow \sigma^2(\theta) \tag{2.8}$$

例 2.4 设 $X_1, \cdots, X_n$ 是来自 $b(1, \theta)$ 的一个样本，θ 的一个估计是 $\overline{X} = \sum X_i/n$. 由中心极限定理

$$\sqrt{n}(\overline{X} - \theta) \xrightarrow{L} N(0, \theta(1 - \theta))$$

对任一参数 $g(\theta)$，若 $g'(\theta)$ 存在，则由定理 1.7 有

$$\sqrt{n}[g(\overline{X}) - g(\theta)] \xrightarrow{L} N(0, [g'(\theta)]^2\theta(1 - \theta))$$

$g(\overline{X})$ 是 $g(\theta)$ 的渐近正态估计. 譬如，若取 $g(\theta) = \theta/(1-\theta)$，则 $g'(\theta) = (1-\theta)^{-2}$，$\overline{X}/(1-\overline{X})$ 是 $\theta/(1-\theta)$ 的渐近正态估计，其渐近方差为 $\frac{\theta}{n(1-\theta)^3}$. 当然，$g(\overline{X})$ 也是 $g(\theta)$ 的相合估计.

例 2.5　设 $X_1,\cdots,X_n$ 是来自 $N(\mu,\sigma^2)$ 的一个样本．记 $Q^2=\sum_{i=1}^n(X_i-\bar{X})^2$，则 $Q^2/\sigma^2\sim\chi^2(n-1)$．由 χ^2 分布的性质可求出

$$\mathrm{E}(Q/\sigma)=\sqrt{2}\,\Gamma\left(\frac{n}{2}\right)\Big/\Gamma\left(\frac{n-1}{2}\right)$$

于是，$\hat{\sigma}=\Gamma\left(\frac{n-1}{2}\right)Q\Big/\left[\sqrt{2}\,\Gamma\left(\frac{n}{2}\right)\right]$ 是 σ 的无偏估计．下面我们证明 $\sqrt{n}(\hat{\sigma}-\sigma)\xrightarrow{L}N(0,\sigma^2/2)$，从而 $\hat{\sigma}$ 是渐近正态的．

因 $\chi^2(n-1)$ 变量可看作 $n-1$ 个独立的 $N(0,1)$ 变量的平方和，故由中心极限定理有 $\sqrt{n-1}\left(\frac{Q^2/\sigma^2}{n-1}-1\right)\xrightarrow{L}N(0,2)$．令 $h(t)=\sqrt{t}$，由定理 1.7 有

$$\sqrt{n-1}\left(\frac{Q/\sigma}{\sqrt{n-1}}-1\right)\xrightarrow{L}N(0,1/2)$$

即

$$Q-\sqrt{n-1}\,\sigma\xrightarrow{L}N(0,\sigma^2/2)$$

再利用 Stirling 公式 $\lim\limits_{x\to\infty}\{\Gamma(x)/[\sqrt{2\pi}x^{x-1/2}\mathrm{e}^{-x}]\}=1$，有

$$\lim_{n\to\infty}\frac{\sqrt{n}\,\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2}\,\Gamma\left(\frac{n}{2}\right)}=\lim_{n\to\infty}\frac{\sqrt{n}\left(\frac{n-1}{2}\right)^{\frac{n}{2}-1}\mathrm{e}^{-\frac{n-1}{2}}}{\sqrt{2}\left(\frac{n}{2}\right)^{\frac{n-1}{2}}\mathrm{e}^{-\frac{n}{2}}}$$

$$=\lim_{n\to\infty}\left\{\mathrm{e}^{\frac{1}{2}}\left(\frac{n-1}{n}\right)^{\frac{n}{2}-1}\right\}$$

$$=1$$

所以

$$\sqrt{n}(\hat{\sigma}-\sigma)=\frac{\sqrt{n}\,\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2}\,\Gamma\left(\frac{n}{2}\right)}(Q-\sqrt{n-1}\,\sigma)+$$

$$\left(\frac{\sqrt{n(n-1)}\,\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2}\,\Gamma\left(\frac{n}{2}\right)}-\sqrt{n}\right)\sigma$$

$$\xrightarrow{L} N(0,\sigma^2/2)$$

渐近正态性是估计的优良性质，但我们必须注意到，估计的渐近正态性只是反映了当 $n\longrightarrow\infty$ 时估计的性质，它并不能说明为达到所需要的精度样本容量必须为多大才行. 事实上，$g(\theta)$ 的渐近正态的估计序列可以有许多，它们之间也存在一个优劣的关系，这就引出了相对渐近效的概念. 先看一个例子.

例 2.6 设 $X_1,\cdots,X_n$ 为来自 $N(\theta,1)$ 的一个样本，我们可以用样本均值 $\overline{X}_n$ 估计 θ，则 $\sqrt{n}(\overline{X}_n-\theta)\xrightarrow{L}N(0,1)$（此处实为 $\overline{X}_n$ 的真实分布）. 另外，由于 θ 是 $N(\theta,1)$ 分布的中位数，我们也可用样本中位数 $m_{0.5}$ 估计 θ，且（见定理 1.10）

$$\sqrt{n}(m_{0.5}-\theta)\xrightarrow{L}N\left(0,\frac{\pi}{2}\right)$$

$\overline{X}_n$ 和 $m_{0.5}$ 都是 θ 的渐近无偏的估计，但它们的渐近方差不同，$\overline{X}$ 的渐近方差为 $\frac{1}{n}$，$m_{0.5}$ 的渐近方差为 $\frac{\pi}{2n}$，由于 $\frac{1}{n}<\frac{\pi}{2n}$，所以前者优于后者. 当 n 大时，对同样的 n，$\overline{X}_n$ 的精度高于 $m_{0.5}$，其比为 $\pi/2$，约为 1.57. 换句话说，要达到同样的精度，用 $m_{0.5}$ 所需的样本容量应为 $\overline{X}$ 的样本容量的 1.57 倍. 比如，要使估计的标准差不大于 0.01，用 $\overline{X}_n$ 须有 10 000 个观测值，而用 $m_{0.5}$ 则须 15 700 个观测值.

定义 2.6 设 $\hat{\theta}_n,\tilde{\theta}_n$ 是 θ 的两个渐近正态估计，其渐近方差分别为 σ_{1n}^2 和 σ_{2n}^2，则称

$$e(\theta,\hat{\theta}_n,\tilde{\theta}_n)=\lim_{n\to\infty}\frac{\sigma_{2n}^2}{\sigma_{1n}^2}$$

为 $\hat{\theta}_n$ 对 $\tilde{\theta}_n$ 的相对渐近效.

在例 2.6 中，$m_{0.5}$ 对 $\overline{X}_n$ 的相对渐近效为 $2/\pi$，约为 0.64.

§2.2 无偏估计

§2.2.1 无偏性

无偏性是统计问题中应用很广的一个准则，其定义及一个例子已

在 2.1 节中给出. 在此,我们先对无偏估计作以下三点说明.

(1)无偏估计不一定存在.

例 2.7　设 $X\sim b(n,\theta)$, $0<\theta<1$, 考虑 $g(\theta)=1/(1-\theta)$, 倘若 $T(X)$ 是 $g(\theta)$ 的无偏估计,则

$$\sum_{i=1}^{n}\binom{n}{i}T(i)\theta^{i}(1-\theta)^{n-i}=g(\theta),\quad \forall\,\theta\in(0,1) \tag{2.9}$$

即

$$\sum_{i=1}^{n}\binom{n}{i}T(i)\theta^{i}(1-\theta)^{n-i+1}-1=0,\quad \forall\,\theta\in(0,1)$$

由于上式左端是 θ 的 $n+1$ 次多项式,无论取什么样的 $T(X)$,它都不可能有无穷个解,因此上式不可能成立. 由此可见,只有当 $g(\theta)$ 为 θ 的 n 次多项式时,才能找到 $T(X)$ 使(2.9)式成立,此种 $g(\theta)$ 才有无偏估计. 而当 $g(\theta)$ 不能表示为 θ 的 n 次多项式时, $g(\theta)$ 的无偏估计不存在.

在统计中,一般将存在无偏估计的参数函数称为可估参数. 今后谈论无偏估计时,总是对可估参数而言,因为对不存在无偏估计的参数去讨论其无偏估计是没有意义的.

(2)对可估参数,无偏估计一般不唯一.

可估参数 $g(\theta)$ 的无偏估计的唯一性等同于 0 的无偏估计的唯一性. 若 $\hat{g}_1(X)$ 和 $\hat{g}_2(X)$ 是 $g(\theta)$ 的两个无偏估计,则 $\hat{g}_1(X)-\hat{g}_2(X)$ 便是 0 的无偏估计,因此, $\hat{g}_1(X)$ 和 $\hat{g}_2(X)$ 是否(几乎处处)相等便取决于 0 的无偏估计是否唯一. 在第一章中我们已指出,当统计结构完备时,0 的无偏估计唯一,反之亦然. 重复抽样结构是不完备的,因此,在重复抽样结构中,可估参数的无偏估计不唯一. 譬如,若 $\hat{g}(X)$ 是 $g(\theta)$ 的无偏估计,则 $\hat{g}(X)+X_1-X_2$ 也是 $g(\theta)$ 的无偏估计.

在统计中,把可估参数 $g(\theta)$ 的无偏估计归为一类,记为 U_g.

(3)无偏估计不一定是好估计.

例 2.8　设 θ 是贝努里试验的成功概率, X 表示首次成功发生前的试验次数,则 X 服从几何分布,其概率函数为

$$p(x;\theta)=\theta(1-\theta)^{x},\quad x=0,1,2,\cdots,\quad 0<\theta<1$$

若 $T(X)$ 为 θ 的无偏估计,则应有

$$\sum_{x=0}^{\infty}T(x)\theta(1-\theta)^x=\theta,\forall\ \theta\in(0,1)$$

由实数稠密性立知 $T(0)=1,T(i)=0,i=1,2,\cdots$,即

$$T(X)=\begin{cases}1, & X=0\\ 0, & X=1,2,\cdots\end{cases}$$

这样,θ 的估计非 0 即 1. 由几何分布的完备性(见习题 1.40),该无偏估计是唯一的. 然而,因为不管样本观测值是多少,这个估计的取值只能是 0 或 1,因此,它是一个很差的估计.

§2.2.2 一致最小方差无偏估计

当样本容量 $n>1$ 时,可估参数的无偏估计不唯一. 设 $g(\theta)$ 为可估参数,我们把 $g(\theta)$ 的所有无偏估计组成的类记为 U_g,如何从 U_g 中选取一个较好的估计是我们关心的问题. 譬如,是否存在这样一个无偏估计,其均方误差(方差)在 U_g 中对 Θ 中所有 θ 一致达到最小呢? 如果这样的无偏估计存在,又该怎样把它找出来呢?下面我们就对这两个问题进行讨论.

定义 2.7 设 $g(\theta)$ 是可估参数,如果 $T(X)$ 是 $g(\theta)$ 的无偏估计,且对 U_g 中任一个估计 $\varphi(X)$,有

$$\mathrm{Var}_\theta(T(X))\leqslant\mathrm{Var}_\theta(\varphi(X)),\quad\forall\ \theta\in\Theta$$

则称 $T(X)$ 为 $g(\theta)$ 的一致最小方差无偏估计(Uniformly Minimum Variance Unbiased Estimate),简记为 UMVUE.

为研究 UMVUE,必须注意到充分统计量的一个重要作用:降低无偏估计的方差.

引理 2.2 设 $S(X)$ 是分布族 $\{p_\theta,\theta\in\Theta\}$ 的充分统计量,$\varphi(X)$ 是 $g(\theta)$ 的无偏估计,令 $T(X)=\mathrm{E}(\varphi(X)|S(X))$,则 $T(X)$ 也是 $g(\theta)$ 的无偏估计,且

$$\mathrm{Var}(T(X))\leqslant\mathrm{Var}(\varphi(X)).$$

证明:因 $S(X)$ 是充分统计量,故 $T(X)=\mathrm{E}(\varphi(X)|S(X))$ 与 θ 无关,$T(X)$ 是统计量. 易见

$$\mathrm{E}(T(X))=\mathrm{E}[\mathrm{E}(\varphi(X)|S(X))]=\mathrm{E}(\varphi(X))=g(\theta)$$

而

$$\begin{aligned}\mathrm{Var}(\varphi(X)) &= \mathrm{E}(\varphi(X)-g(\theta))^2\\ &= \mathrm{E}(\varphi(X)-T(X)+T(X)-g(\theta))^2\\ &= \mathrm{E}(\varphi(X)-T(X))^2+\mathrm{E}(T(X)-g(\theta))^2\\ &\geqslant \mathrm{Var}(T(X))\end{aligned}$$

其中交叉乘积项

$$\begin{aligned}&\mathrm{E}[(\varphi(X)-T(X))(T(X)-g(\theta)]\\ =&\mathrm{E}\{\mathrm{E}[(\varphi(X)-T(X))(T(X)-g(\theta))\mid S(X)]\}\\ =&\mathrm{E}\{(T(X)-g(\theta))\mathrm{E}[\varphi(X)-T(X)\mid S(X)]\}\\ =&0\end{aligned}$$

引理得证.

由证明可以看出，只要 $\mathrm{Var}_\theta(\varphi(X))<\infty$，则 $\mathrm{Var}_\theta(T(X))=\mathrm{Var}_\theta(\varphi(X))$ 的充要条件是 $T(x)=\varphi(x)$，a.s.，由此可见，恰当地使用充分统计量可以降低无偏估计的方差. 问题是，什么样的充分统计量可以最大程度地降低方差，另外，这样得到的无偏估计是否是 UMVUE 呢？下面的定理回答了这两个问题.

定理 2.3 设 $S(X)$ 为 $\{P_\theta,\theta\in\Theta\}$ 的完备充分统计量，$g(\theta)$ 为可估参数，则 $g(\theta)$ 的 UMVUE 存在，它是 $S(X)$ 的函数且在几乎处处意义下是唯一的.

证明：因 $g(\theta)$ 是可估参数，故 U_g 非空，在 U_g 中取定 $\varphi(X)$，令 $T(X)=\mathrm{E}(\varphi(X)\mid S(X))$，则 $T(X)$ 是 $S(X)$ 的函数. 由引理 2.2 知，$T(X)\in U_g$，且 $\mathrm{Var}_\theta(T(X))\leqslant\mathrm{Var}_\theta(\varphi(X))$.

任取 $\hat{g}(X)\in U_g$，令 $T^*(x)=\mathrm{E}(\hat{g}(X)\mid S(x))$，则

$$\mathrm{E}(T^*(X)-T(X))=0,\quad \forall\,\theta\in\Theta$$

而 $T^*(x)$，$T(x)$ 均是 $S(x)$ 的函数，由 $S(x)$ 的完备性知

$$T^*(x)=T(x),\text{a.s.}$$

此即

$$\mathrm{Var}(T(X))\leqslant\mathrm{Var}(\hat{g}(X)),\quad \forall\,\theta\in\Theta$$

由此，从 U_g 中任一估计出发均可得到一个相同的 $T(X)$，该 $T(X)$ 是 $S(X)$ 的函数，它是 $g(\theta)$ 的（几乎处处）唯一的 UMVUE. 证毕.

§2.2.3 例题

由上述讨论，只要完备充分统计量存在，可估参数的 UMVUE 一定存在. 定理 2.3 及其证明提供了两种求 UMVUE 的方法.

方法 1：寻找完备充分统计量的函数使之成为 $g(\theta)$ 的无偏估计；

方法 2：任取 $g(\theta)$ 的一个无偏估计并将之对完备充分统计量求条件期望.

例 2.9 设 $X_1,\cdots,X_n$ 是来自 $b(1,\theta)$，$0<\theta<1$，的一个样本，由指数型分布族的性质知，$S(X)=\sum X_i$ 是完备充分统计量.

(1) 对 θ，因为 $E(S(X))=n\theta$，因而，$\overline{X}=S(X)/n$ 是 θ 的无偏估计，从而也是 θ 的 UMVUE.

(2) 对 $g(\theta)=\theta^k+(1-\theta)^{n-k}$（$k$ 为整数，$0\leqslant k\leqslant n$）. 要直接找一个 $S(X)$ 的函数 $h(S(X))$ 使之成为 $g(\theta)$ 的无偏估计是很困难的，但我们可以方便地找到 $g(\theta)$ 的一个无偏估计. 令

$$\varphi_1(X)=\begin{cases}1, & \sum_{i=1}^k X_i=k\\ 0, & \text{其它}\end{cases}$$

$$\varphi_2(X)=\begin{cases}1, & \sum_{i=k+1}^n X_i=0\\ 0, & \text{其它}\end{cases}$$

令 $\varphi(X)=\varphi_1(X)+\varphi_2(X)$，则 $E(\varphi(X))=Pr\left\{\sum_{i=1}^k X_i=k\right\}+Pr\left\{\sum_{i=k+1}^n X_i=0\right\}=Pr\{X_i=1,i=1,\cdots,k\}+Pr\{X_j=0,j=k+1,\cdots,n\}=\theta^k+(1-\theta)^{n-k}$，可见 $\varphi(X)$ 是 $g(\theta)$ 的一个无偏估计，由定理 2.3 知 $E(\varphi(X)\mid S(X))$ 是 $g(\theta)$ 的 UMVUE. 余下只要算出 $T(X)=E(\varphi(X)\mid S(X))$ 即可.

记 $S(X)$ 取值为 s，当 $k\leqslant s$ 时，

$$\begin{aligned}E(\varphi_1(X)\mid S(X)=s)&=Pr(\varphi_1(X)=1\mid S(X)=s)\\&=\frac{Pr\left\{\sum_{i=1}^k X_i=k,\sum_{i=k+1}^n X_i=s-k\right\}}{Pr(S(X)=s)}\\&=\frac{\theta^k\cdot\binom{n-k}{s-k}\theta^{s-k}(1-\theta)^{n-s}}{\binom{n}{s}\theta^s(1-\theta)^{n-s}}\end{aligned}$$

$$= \binom{n-k}{s-k} \Big/ \binom{n}{s}$$

当 $k>s$ 时，$\mathrm{E}(\varphi_1(X)|S(X)=s)=0$. 同理，$k\geqslant s$ 时，

$$\mathrm{E}(\varphi_2(X)|S(X)) = \binom{k}{s} \Big/ \binom{n}{s}$$

而 $k<s$ 时，$\mathrm{E}(\varphi_2(X)|S(X)=s)=0$. 因此，$g(\theta)$的 UMVUE 为

$$T(X) = \begin{cases} \binom{n-k}{s-k} \Big/ \binom{n}{s}, & k<s \\ \binom{k}{s} \Big/ \binom{n}{s}, & k>s \\ 2 \Big/ \binom{n}{s}, & k=s \end{cases}$$

例 2.10　某厂生产的产品其废品率为 θ，现将该产品包装成盒，每盒抽 n 个产品逐个检验，得废品数 X（假设盒中产品数远远大于 n，可认为 $X\sim b(n,\theta)$）. 当 $x\leqslant 2$ 时，商店接收该盒产品，$x\geqslant 3$ 时拒收. 令

$$g(\theta) = (1-\theta)^n + n\theta(1-\theta)^{n-1} + n(n-1)\theta^2(1-\theta)^{n-2}$$

则 $g(\theta)$是通过概率. 厂方自然很关心 $g(\theta)$的估计.

现设抽了 r 盒产品进行检验，第 i 盒查得废品数为 X_i，$i=1,\cdots,r$，则 $X_i\sim b(n,\theta)$，$S(X)=\sum_{i=1}^{r} X_i$ 为完备充分统计量，且 $S(X)\sim b(rn,\theta)$. 令

$$\varphi(X) = \begin{cases} 1 & \text{第一盒被接收}(x_1\leqslant 2) \\ 0 & \text{其它} \end{cases}$$

则

$$\mathrm{E}_\theta\varphi(X) = P_\theta(\varphi(X)=1) = g(\theta)$$

为计算 $\mathrm{E}(\varphi(X)|S(X))$，令 B_i 表示“在第一盒中有且只有 i 个废品”，$i=0,1,2$. 则当 B_0,B_1,B_2 中有一个发生时，第一盒被接收. 由于 B_0，B_1，B_2 不相容，记 $B=\bigcup_{i=0}^{2}B_i$，则 $\varphi(x)=I_B$，从而

$$\begin{aligned} &\mathrm{E}(\varphi(X)|S(X)) \\ &= Pr(B|S(X)) = \sum_{i=0}^{2} Pr(B_i|S(X)) \end{aligned}$$

$$=\frac{Pr(X_1=0,S(X)=s)+Pr(X_1=1,S(X)=s)+Pr(X_1=2,S(X)=s)}{Pr(S(X)=s)}$$

$$=\left\{\binom{rn-n}{s}+n\binom{rn-n}{s-1}+n(n-1)\binom{rn-n}{s-2}\right\}\Big/\binom{rn}{s}$$

这就是 $g(\theta)$ 的 UMVUE.

例 2.11 假设某种产品的寿命 X 服从带未知参数 λ 的指数分布，对某固定时间 t_0，我们要估计产品在 t_0 前失效的概率 $P_\lambda(X\leqslant t_0)=1-e^{-\lambda t_0}\triangleq g(\lambda)$.

为此，我们取 n 个这样的产品进行试验，记录其寿命分别为 $X_i,i=1,\cdots,n$. 由指数型分布族性质知 $S(X)=\sum X_i$ 是完备充分统计量. 取 $\varphi(X)=I_{(X_1\leqslant t_0)}$，容易看出，$\varphi(X)$ 是 $g(\lambda)$ 的无偏估计. 在 $S(X)=s$ 的条件下，

$$\begin{aligned}\mathrm{E}(\varphi(X)\mid S(X)=s)&=Pr(X_1\leqslant t_0\mid S(X)=s)\\&=Pr\left(\frac{X_1}{s}\leqslant\frac{t_0}{s}\mid S(X)=s\right)\end{aligned}$$

注意到诸 X_i 独立同分布，服从 $\mathrm{Exp}(\lambda)=Ga(1,\lambda)$，由 Gamma 分布性质（习题 1.8）知 $\frac{X_1}{S(X)}$ 与 $S(X)$ 独立，且 $\frac{X_1}{S(X)}\sim Be(1,n-1)$，故当 $t_0/s\leqslant 1$ 时

$$\begin{aligned}\mathrm{E}(\varphi(X)\mid S(X)=s)&=\int_0^{t_0/s}(n-1)(1-u)^{n-2}\mathrm{d}u\\&=1-(1-t_0/s)^{n-1}\end{aligned}$$

而 $t_0/s>1$ 时，$\mathrm{E}(\varphi(X)\mid S(X)=s)=1$，由此给出 $g(\lambda)$ 的 UMVUE 为

$$T(x)=\begin{cases}1-\left(1-\dfrac{t_0}{S(x)}\right)^{n-1}, & \text{如果 } S(x)\geqslant t_0\\ 1, & \text{如果 } S(x)<t_0\end{cases}$$

例 2.12 设 $X_1,\cdots,X_n$ 是来自 $U(0,\theta)$ 的一个样本，由因子分解定理知 $X_{(n)}=\max(X_1,\cdots,X_n)$ 为充分统计量，其分布密度函数为

$$p(t;\theta)=nt^{n-1}/\theta^n,\ 0<t<\theta.$$

易验证该分布族是完备的，因而 $X_{(n)}$ 也是 θ 的完备统计量. 因 $\mathrm{E}X_{(n)}=\int_0^\theta nt^n\mathrm{d}t/\theta^n=\frac{n}{n+1}\theta$，故 $\frac{n+1}{n}X_{(n)}$ 是 θ 的 UMVUE.

此题还有一种不常见的解法：在已知 $X_{(n)}$ 是 θ 的完备充分统计量后，记 $X_{(n)}$ 取值为 t，设 $\hat{\theta}(t)$ 是 θ 的无偏估计（从而即为 θ 的 UMVUE），则

$$\int_0^\theta \hat{\theta}(t) n t^{n-1}/\theta^n \mathrm{d}t = \theta. \quad \forall\ \theta > 0$$

此即

$$\int_0^\theta \hat{\theta}(t) n t^{n-1} \mathrm{d}t = \theta^{n+1} = \int_0^\theta (n+1) t^n \mathrm{d}t, \quad \forall\ \theta > 0$$

两边求导，立有

$$\hat{\theta}(t) n t^{n-1} = (n+1) t^n, \mathrm{a.s.}$$

从而 $\hat{\theta}(t) = \frac{n+1}{n} t$ 是 θ 它的 UMVUE. 该方法不是普遍适用的，但在能够求出时是比较简单的.

例 2.13 指数分布是可靠性统计中一类很重要的分布，许多产品的寿命服从指数分布. 在可靠性试验中，由于时间等因素的影响，做完全部试验往往很困难，因此截尾试验经常被采用. 假设有 n 个产品投入试验，当第 r 个 $(r<n)$ 产品失效发生时试验停止. 于是，我们能观测到 n 个产品中前 r 个失效产品的寿命，记为 $X_{(1)} \leqslant \cdots \leqslant X_{(r)}$.

假设产品寿命的密度函数为 $p(x;\theta) = \frac{1}{\theta} \mathrm{e}^{-x/\theta}, x>0, \theta>0$. 由因子分解定理知 $S = \sum_{i=1}^r X_{(i)} + (n-r) X_{(r)}$ 是 θ 的充分统计量. 事实上，因为 $S \sim Ga(r, 1/\theta)$，故 S 还是 θ 的完备统计量. 又由于 $ES = r\theta$，所以 $\hat{\theta} = S/r$ 是 θ 的 UMVUE.

更一般地，假设产品的寿命分布为 $p(x;\alpha,\theta) = \frac{1}{\theta} \mathrm{e}^{-(x-\alpha)/\theta}, x>\alpha$，则 $X_{(1)}, \cdots, X_{(r)}$ 的联合密度函数可以写出为

$$\frac{n!}{(n-r)!} \prod_{i=1}^r p(x_{(i)};\alpha,\theta) [1 - F(x_{(r)};\alpha,\theta)]^{n-r}$$

$$= \frac{n!}{(n-r)!} \theta^{-r} \mathrm{e}^{n\alpha/\theta} \exp\{-(nT_1 + T_2)/\theta\} I_{(T_1 > \alpha)} I_{(T_2 > 0)},$$

其中 $T_1 = X_{(1)}$, $T_2 = \sum_{i=2}^r (n-i+1)(X_{(i)} - X_{(i-1)})$. 于是 (T_1, T_2) 是 (α, θ) 的充分统计量. 作变换 $Y_1 = n(X_{(1)} - \alpha)$, $Y_i = (n-i+1) \cdot$

$(X_{(i)}-X_{(i-1)})$，$i=2,\cdots,r$，可证得诸 Y_i 独立同分布，且 $Y_1\sim Exp(1/\theta)$，从而可求出 (T_1,T_2) 的分布并证明 (T_1,T_2) 也是 (α,θ) 的完备统计量，且 $ET_1=\alpha+\theta/n$，$ET_2=(r-1)\theta$，由此即知 $\hat{\alpha}=T_1-T_2/[n(r-1)]$，$\hat{\theta}=T_2/(r-1)$ 是各自参数的 UMVUE.

由上述讨论可见，存在完备充分统计量是可估参数存在 UMVUE 的充分条件，那末这个条件是否是必要条件呢？回答是否定的. 事实上，关于无偏估计是否是 UMVUE 的判断有一个充要条件，见习题2.18，可参阅文献[3].

§2.2.4 U 统计量

上述讨论同样可以应用到非参数统计结构. 在非参数统计结构中寻找 UMVUE，定理 2.3 也是适合的，此处，U 统计量扮演了重要角色.

设统计结构是 $(\mathscr{X},\mathscr{B},\mathscr{P})$，我们要估计参数 $g(P)$. 我们把使 $g(P)$ 能估计出来的最小样本容量称为 $g(P)$ 的阶，记为 m，譬如，若 $g(P)=E_PX$，则 $m=1$，若 $g(P)=\mathrm{Var}_P(X)$，则 $m=2$. 由容量为 m 的样本可给出 $g(P)$ 的无偏估计，称为核，若该无偏估计是样本的对称函数，则称之为对称核. 比如，若 $g(P)=E_PX$，则 X_1 是 $g(P)$ 的对称核；对 $g(P)=\mathrm{Var}_P(X)$，$X_1^2-X_1X_2$ 是核，但不是对称核，然而我们可由核构造对称核：只要把 $m!$ 个可能不同的核加以平均即可. 此处 $\frac{1}{2}[(X_1^2-X_1X_2)+(X_2^2-X_1X_2)]=\frac{1}{2}(X_1-X_2)^2$ 为对称核. 我们用 $f_m(X_1,\cdots,X_m)$ 表示对称核.

现设我们有 n 个样品组成的样本 $(n\geqslant m)$，其中任意 m 个样品都可以给出一个对称核，我们把所有 $\binom{n}{m}$ 个对称核的平均称为 U 统计量，记为 U_n.

$$U_n=\binom{n}{m}^{-1}\sum_{1\leqslant i_1<\cdots<i_m\leqslant n}f_m(X_{i_1},\cdots,X_{i_m}) \tag{2.10}$$

显然，U_n 是 $g(P)$ 的无偏估计，而且 U_n 是样本的对称函数，亦即是说，

U_n 是次序统计量$(X_{(1)},\cdots,X_{(n)})$的函数.对常见的非参数统计结构,$(X_{(1)},\cdots,X_{(n)})$大多是完备充分统计量[4],因而 U_n 即为 $g(P)$ 的 UMVUE.

例 2.14 (1) 设 $\mathscr{P}=\{$所有一阶矩存在的一维分布$\}$,对此分布族,$(X_{(1)},\cdots,X_{(n)})$是完备充分统计量.令 $g(P)$为总体均值,其 $m=1$,对称核为 $f_1(X_1)=X_1$,故由(2.10)知

$$U_n=\sum X_i/n=\overline{X}$$

是总体均值的 UMVUE.

(2) 设 $\mathscr{P}=\{$所有二阶矩存在的一维分布$\}$,则$(X_{(1)},\cdots,X_{(n)})$也是完备充分的.令 $g(P)$为总体方差,其 $m=2$,对称核为 $f_2(X_1,X_2)=\frac{1}{2}(X_1-X_2)^2$,故由(2.10)有

$$\begin{aligned}U_n&=\frac{2}{n(n-1)}\sum_{i<j}\frac{1}{2}(X_i-X_j)^2\\&=\frac{1}{n(n-1)}\Big[(n-1)\sum X_i^2-2\sum_{i<j}X_iX_j\Big]\\&=\frac{1}{n(n-1)}\Big[n\sum X_i^2-\Big(\sum X_i\Big)^2\Big]\\&=\frac{1}{n-1}\sum(X_i-\overline{X})^2=s^2\end{aligned}$$

由此可见,对多数非参数统计问题,样本均值 $\overline{X}$ 和样本方差 s^2 分别是总体均值和方差的 UMVUE.

U 统计量具有很好的大样本性质,比如强相合性和渐近正态性,这使得 U 统计量在非参数统计推断中起到很大作用,在第三章中我们将作进一步讨论.

§2.3 信息不等式

§2.3.1 Fisher 信息量

Fisher 信息量与信息不等式是统计学中两个重要结果,这里先介

绍 Fisher 信息量的概念及其性质.

定义 2.8 设统计结构$(\mathscr{X},\mathscr{B},\{P_\theta,\theta\in\Theta\})$可控,$\Theta$是$\mathbf{R}^k$的子集合. 假如定义在$(\mathscr{X},\mathscr{B})$上取值于$(\mathbf{R}^k,\mathscr{B}_{\mathbf{R}^k})$的随机向量

$$\boldsymbol{S}_\theta(X)=\left(\frac{\partial\ln p_\theta(x)}{\partial\theta_1},\cdots,\frac{\partial\ln p_\theta(x)}{\partial\theta_k}\right)'$$

满足

(1) $S_\theta(x)$对一切$\theta\in\Theta$有定义;

(2) $\mathrm{E}_\theta\boldsymbol{S}_\theta(X)=0,\forall\ \theta\in\Theta$;

(3) $\boldsymbol{S}_\theta(X)$模平方可积,即$\mathrm{E}_\theta\|\boldsymbol{S}_\theta(X)\|^2<\infty$.

则把$\boldsymbol{S}_\theta(X)$的协差阵

$$\boldsymbol{I}(\theta)=\mathrm{Var}_\theta(\boldsymbol{S}_\theta(X))=\mathrm{E}_\theta[\boldsymbol{S}_\theta(X)\boldsymbol{S}'_\theta(X)]\tag{2.11}$$

称为该统计结构的 Fisher 信息矩阵,简称 Fisher 信息,$k=1$时$\boldsymbol{I}(\theta)$常称为 Fisher 信息量.

关于 Fisher 信息,首先有一个存在性问题. 对此,众知的结论是:Cramer-Rao 正则族中 Fisher 信息存在.

定义 2.9 分布族$\{p_\theta(x),\theta\in\Theta\}$称为 Cramer-Rao 正则族,如果

(1) Θ是$\mathbf{R}^k$上开矩形;

(2) $\partial\ln p_\theta(x)/\partial\theta_i,i=1,\cdots,k$,对所有$\theta\in\Theta$都存在;

(3) 支撑$A=\{x: p_\theta(x)>0\}$与θ无关;

(4) 对$p_\theta(x)$,积分与微分可交换;

(5) 对一切$1\leqslant i,j\leqslant k,\forall\ \theta\in\Theta,\mathrm{E}_\theta\left|\frac{\partial\ln p_\theta(x)}{\partial\theta_i}\frac{\partial\ln p_\theta(x)}{\partial\theta_j}\right|<\infty$.

易见,定义 2.9 中的(2),(5)分别与定义 2.8 中的(1),(3)等价,而定义 2.9 中的(3)加(4)可推出定义 2.8 中的(2),因此,Cramer-Rao 正则族存在 Fisher 信息. 回顾指数族的性质,其自然参数空间为凸集,一般而言,指数族为 Cramer-Rao 正则族.

$k=1$和$k=2$是两种最常见的情形. 看两个例子.

例 2.15($k=1$) 易知 Poisson 分布族$\{P(\lambda),\lambda>0\}$为 Cramer-Rao 正则族,其 Fisher 信息存在,易计算

$$S_\lambda(x)=\partial\ln p_\lambda(x)/\partial\lambda=x/\lambda-1$$

$$I(\lambda)=E_\lambda(S_\lambda(X))^2=\mathrm{Var}_\lambda(X/\lambda)=1/\lambda$$

例 2.16（$k=2$）　正态分布族$\{N(\mu,\sigma^2)(\mu,\sigma^2)\in\mathbf{R}\times\mathbf{R}^+\}$为Cramer-Rao 正则族，其 Fisher 信息存在. 记 $\theta=(\mu,\sigma^2)$，则

$$\boldsymbol{S}_\theta(x)=\left\{\frac{x-\mu}{\sigma^2},\frac{(x-\mu)^2}{2\sigma^4}-\frac{1}{2\sigma^2}\right\}'$$

$$\boldsymbol{I}(\theta)=\mathrm{Var}_\theta(\boldsymbol{S}_\theta(X))=(I_{ij})_{2\times 2}$$

其中

$$I_{11}=\mathrm{Var}_\theta\left(\frac{x-\mu}{\sigma^2}\right)=1/\sigma^2$$

$$I_{22}=\mathrm{Var}_\theta\left(\frac{(x-\mu)^2}{2\sigma^4}\right)=1/(2\sigma^4)$$

$$I_{12}=E_\theta\left\{\frac{x-\mu}{\sigma^2}\left[\frac{(x-\mu)^2}{2\sigma^4}-\frac{1}{2\sigma^2}\right]\right\}=0$$

如果进一步假设 $\ln p_\theta(x)$可对 θ 求二阶偏导，且积分与微分可交换次序（这在多数场合是成立的，如指数族），则 Fisher 信息的计算可如下进行.

记 $S_\theta^{(i)}=\dfrac{\partial\ln p_\theta(x)}{\partial\theta_i}$，$i=1,\cdots,k$，则 $\boldsymbol{I}(\theta)$的(i,j)元素为

$$I_{ij}=E(S_\theta^{(i)}S_\theta^{(j)})$$

又由于 $ES_\theta^{(i)}=0$，故

$$\begin{aligned}0=\frac{\partial ES_\theta^{(i)}}{\partial\theta_j}&=\int\frac{\partial\{S_\theta^{(i)}p_\theta(x)\}}{\partial\theta_j}\mathrm{d}x\\&=\int\frac{\partial S_\theta^{(i)}}{\partial\theta_j}p_\theta(x)\mathrm{d}x+\int S_\theta^{(i)}\frac{\partial p_\theta(x)}{\partial\theta_j}\mathrm{d}x\\&=\int\frac{\partial^2\ln p_\theta(x)}{\partial\theta_i\partial\theta_j}p_\theta(x)\mathrm{d}x+\int S_\theta^{(i)}\frac{\partial\ln p_\theta(x)}{\partial\theta_j}p_\theta(x)\mathrm{d}x\\&=E_\theta\left\{\frac{\partial^2\ln p_\theta(x)}{\partial\theta_i\partial\theta_j}\right\}+E[S_\theta^{(i)}S_\theta^{(j)}]\end{aligned}$$

由此即有

$$I_{ij}=-E_\theta\left\{\frac{\partial^2\ln p_\theta(x)}{\partial\theta_i\partial\theta_j}\right\}\tag{2.12}$$

在计算 Fisher 信息时，(2.12)式通常比(2.11)式简单得多. 主要

是因为(2.12)式涉及的求期望的阶数要低. 比如,在例 2.15 中,用(2.12)有 $I(\lambda)=-\mathrm{E}_\lambda\{\partial^2\ln p_\lambda/\partial\lambda^2\}=\mathrm{E}_\lambda(X/\lambda^2)=1/\lambda$;在例 2.16 中,

$$I_{11}=-\mathrm{E}_\theta\left\{\frac{\partial^2\ln p_\theta(x)}{\partial\mu^2}\right\}=\frac{1}{\sigma^2}$$

$$I_{22}=-\mathrm{E}_\theta\left\{\frac{\partial^2\ln p_\theta(x)}{\partial(\sigma^2)^2}\right\}=\mathrm{E}\left\{\frac{(x-\mu)^2}{\sigma^6}\right\}-\frac{1}{2\sigma^4}=\frac{1}{2\sigma^4}$$

$$I_{12}=-\mathrm{E}_\theta\left\{\frac{\partial^2\ln p_\theta(x)}{\partial\mu\partial\sigma^2}\right\}=\mathrm{E}\left(\frac{x-\mu}{\sigma^4}\right)=0$$

对重复抽样结构,Fisher 信息的计算可简化. 设 $X_1,\cdots,X_n$ 独立同分布,$X=(X_1,\cdots,X_n)'$,则

$$\boldsymbol{S}_\theta(X)=\frac{\partial}{\partial\theta}\ln\prod_{i=1}^n p_\theta(X_i)=\sum\boldsymbol{S}_\theta(X_i)$$

于是,在 $\boldsymbol{I}(\theta)$ 存在时,

$$\boldsymbol{I}_n(\theta)=\mathrm{Var}_\theta(\boldsymbol{S}_\theta(X))=\sum_{i=1}^n\mathrm{Var}_\theta(\boldsymbol{S}_\theta(X_i))=n\boldsymbol{I}_1(\theta)$$

其中,$I_n(\theta)$ 表示 n 个观测值的 Fisher 信息. 事实上,由上述过程可知,由独立变量组成的向量的 Fisher 信息,不论其组成部分是否同分布,都等于各组成部分 Fisher 信息之和. 该性质可简化 $\boldsymbol{I}(\theta)$ 的计算.

§2.3.2 Fisher 信息与充分统计量

对统计量,我们可以用其诱导统计结构的 Fisher 信息定义该统计量的 Fisher 信息.

定义 2.10 设 $T(X)$ 是统计结构 $(\mathscr{X},\mathscr{B},\{P_\theta,\theta\in\Theta\})$ 上的统计量,$(\mathscr{T},\mathscr{C},\{P_\theta^T,\theta\in\Theta\})$ 是 $T(X)$ 的诱导统计结构. 如果 $(\mathscr{T},\mathscr{C},\{P_\theta^T,\theta\in\Theta\})$ 上的 Fisher 信息存在,则称其为统计量 $T(X)$ 的 Fisher 信息,记为 $I_T(\theta)$.

例 2.17 设 $X_1,\cdots,X_n$ 是来自 $N(\mu,\sigma^2)$ 的一个样本,这是一个重复抽样结构,记 $\theta=(\mu,\sigma^2)$,其 Fisher 信息为(参见例 2.16)

$$\boldsymbol{I}(\theta)=\begin{pmatrix}n/\sigma^2 & 0\\ 0 & n/(2\sigma^4)\end{pmatrix}$$

定义统计量 $\boldsymbol{T}(X)=(T_1,T_2)=(\overline{X},s^2)$,则 T_1 与 T_2 独立,且

$T_1 \sim N(\mu, \sigma^2/n)$，$T_2 \sim Ga\left(\frac{n-1}{2}, \frac{n-1}{2\sigma^2}\right)$. 故$(T_1, T_2)$的联合密度函数为

$$p^T(t_1, t_2; \theta) = \sqrt{\frac{n}{2\pi\sigma^2}} e^{-\frac{n(t_1-\mu)^2}{2\sigma^2}} \frac{1}{\Gamma\left(\frac{n-1}{2}\right)} \left(\frac{n-1}{2\sigma^2}\right)^{\frac{n-1}{2}} t_2^{\frac{n-1}{2}-1} e^{-\frac{n-1}{2\sigma^2}t_2}$$

于是

$$\ln p^T(t_1, t_2; \theta) = -\frac{1}{2}\ln\sigma^2 - \frac{n(t_1-\mu)^2}{2\sigma^2} - \frac{n-1}{2}\ln\sigma^2 - \frac{n-1}{2\sigma^2}t_2 + c(t_1, t_2),$$

其中 $c(t_1, t_2) = \frac{1}{2}\ln n - \frac{1}{2}\ln(2\pi) - \ln\Gamma\left(\frac{n-1}{2}\right) + \frac{n-3}{2}\ln t_2 + \frac{n-1}{2}\ln\frac{n-1}{2}$ 与 μ 和 σ^2 无关，

$$\frac{\partial \ln p^T}{\partial \mu} = \frac{n(t_1-\mu)}{\sigma^2}$$

$$\frac{\partial \ln p^T}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{n(t_1-\mu)^2 + (n-1)t_2}{2\sigma^4}$$

$$\frac{\partial^2 \ln p^T}{\partial \mu^2} = -\frac{n}{\sigma^2}$$

$$\frac{\partial^2 \ln p^T}{\partial \mu \partial \sigma^2} = -\frac{n(t_1-\mu)}{\sigma^4}$$

$$\frac{\partial^2 \ln p^T}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{n(t_1-\mu)^2 + (n-1)t_2}{\sigma^6}$$

注意到 $E(T_1) = \mu$，$E(T_1-\mu)^2 = \sigma^2/n$，$E(T_2) = \sigma^2$，由(2.12)可算出

$$I_T(\theta) = \begin{pmatrix} n/\sigma^2 & 0 \\ 0 & n/(2\sigma^4) \end{pmatrix}$$

本例中，$I_T(\theta)$与$I(\theta)$是一致的，这不是孤立的现象，这是充分统计量的一个重要性质. 下面我们介绍统计量的 Fisher 信息的二个性质.

定理 2.4 设$\{P_\theta, \theta \in \Theta\}$是 Cramer-Rao 正则族，其 Fisher 信息存在，记为 $I(\theta)$. 又设 $T(X)$是该统计结构上的统计量，其 Fisher 信息存在，记为 $I_T(\theta)$. 则

$$I_T(\theta) \leqslant I(\theta)$$

证明：记

$$S_\theta(X) = \frac{\partial}{\partial \theta}\ln p(x;\theta); \quad S_\theta^T(t) = \frac{\partial}{\partial \theta}\ln p^T(t;\theta)$$

则

$$I(\theta) = \mathrm{E}(S_\theta S_\theta'); \quad I_T(\theta) = \mathrm{E}(S_\theta^T S_\theta^{T\prime})$$

对任意的 $B \in \mathscr{B}_{\mathscr{T}}$，有

$$\begin{aligned}
\int_B \frac{\partial}{\partial \theta}\ln p_\theta^T(t)\mathrm{d}P_\theta^T(t) &= \int_B \frac{\partial}{\partial \theta} p_\theta^T(t)\mathrm{d}\mu^T \\
&= \frac{\partial}{\partial \theta}\int_B p^T(t)\mathrm{d}\mu^T = \frac{\partial}{\partial \theta} p_\theta^T(B) \\
&= \frac{\partial}{\partial \theta} P(T^{-1}(B)) = \frac{\partial}{\partial \theta}\int_{T^{-1}(B)} p_\theta(x)\mathrm{d}\mu \\
&= \int_{T^{-1}(B)} \frac{\partial}{\partial \theta} p_\theta(x)\mathrm{d}\mu \\
&= \int_{T^{-1}(B)} \frac{\partial}{\partial \theta}\ln p_\theta(x)\mathrm{d}P_\theta(x)
\end{aligned}$$

因 B 是任意的，由条件期望的定义有

$$S_\theta^T(t) = \mathrm{E}_\theta(S_\theta(X)|t), \quad \text{a.s.}$$

令 $W = S_\theta(x) - S_\theta^T(T(x))$，则

$$\begin{aligned}
\mathrm{E}(WW') &= \mathrm{E}[(S_\theta - S_\theta^T)(S_\theta - S_\theta^T)'] \\
&= \mathrm{E}(S_\theta S_\theta') - \mathrm{E}(S_\theta^T S_\theta') - \mathrm{E}(S_\theta S_\theta^{T\prime}) + \mathrm{E}(S_\theta^T S_\theta^{T\prime}),
\end{aligned}$$

利用 $\mathrm{E}(S_\theta S_\theta^{T\prime}) = \mathrm{E}[\mathrm{E}(S_\theta S_\theta^{T\prime}|T)] = \mathrm{E}(S_\theta^T S_\theta^{T\prime})$，立有

$$\mathrm{E}(WW') = \mathrm{E}(S_\theta S_\theta') - \mathrm{E}(S_\theta^T S_\theta^{T\prime}) \geqslant 0$$

证毕.

从直观上看，定理 2.4 的结论也是显然的，因为统计量是对样本的加工，它不可能比样本自身提供更多的信息.

定理 2.5 在定理 2.4 的条件下，$I_T(\theta) = I(\theta)$，$\forall\ \theta \in \Theta$ 的充要条件是 $T(X)$ 是充分统计量.

证明：必要性：由定理 2.4 的证明过程及 $I_T(\theta) = I(\theta)$ 知

$$I(\theta) - I_T(\theta) = \mathrm{E}[(S_\theta - S_\theta^T)(S_\theta - S_\theta^T)'] = 0$$

从而

$$S_\theta(x) = S_\theta^T(T(x)), \quad \text{a.s.}$$

记 $d\boldsymbol{\theta} = (d\theta_1, \cdots, d\theta_k)'$，任取 $\boldsymbol{\theta}_0 \in \Theta$，将全微分 $(\boldsymbol{S}_\theta(x))'d\boldsymbol{\theta} = (\boldsymbol{S}_\theta^T(T(x))'d\boldsymbol{\theta}$从 θ_0 到 $\boldsymbol{\theta}$ 积分，给出

$$\ln p_\theta(x) - \ln p_{\theta_0}(x) = \ln p_\theta^T(T(x)) - \ln p_{\theta_0}^T(T(x))$$

此即

$$p_\theta(x) = g(\boldsymbol{\theta}, T(x))h(x)$$

其中 $g(\boldsymbol{\theta}, T(x)) = p_\theta^T(T(x))$，$h(x) = p_{\theta_0}(x)/p_{\theta_0}^T(T(x))$. 由因子分解定理，$T(X)$是充分统计量.

充分性的证明可查阅文献[3].

定理 2.5 表明，当且仅当 $T(X)$是充分统计量时，$T(X)$所含 $\boldsymbol{\theta}$ 的信息与样本一致，这说明，Fisher 将 $\boldsymbol{I}(\theta)$称为信息量确有一定的根据.

§2.3.3　信息不等式

信息不等式也称 Cramer-Rao 不等式，它是用 Fisher 信息表示无偏估计的方差(协差阵)下限的一个不等式.

定理 2.6　设 $\{p_\theta(x), \theta \in \Theta\}$ 为 Cramer-Rao 正则族，$\Theta \subset \mathbf{R}^k$，其 Fisher 信息 $\boldsymbol{I}(\theta)$是非奇矩阵，并设 $\boldsymbol{g}(\theta) = (g_1(\theta), \cdots, g_s(\theta))'$，$s \leqslant k$，且 $\partial g_i(\theta)/\partial\theta_j$ 对一切 $i = 1, \cdots, s, j = 1, \cdots, k$ 都存在. 假设 $\boldsymbol{T}(X)$是 $\boldsymbol{g}(\theta)$的模平方可积的无偏估计，记 $\boldsymbol{\Delta} = \mathrm{E}(\boldsymbol{T}\boldsymbol{S}_\theta') = \dfrac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\boldsymbol{g}(\theta)$. 则有

$$\mathrm{Var}_\theta(\boldsymbol{T}(X)) \geqslant \boldsymbol{\Delta}\boldsymbol{I}^{-1}(\theta)\boldsymbol{\Delta}', \quad \forall\, \boldsymbol{\theta} \in \Theta \tag{2.13}$$

其中 $\boldsymbol{\Delta}\boldsymbol{I}^{-1}(\theta)\boldsymbol{\Delta}'$称为 $\boldsymbol{g}(\theta)$的无偏估计协差阵的下界，或称为 $\boldsymbol{g}(\theta)$的无偏估计的 Cramer-Rao 下界，简称 C-R 下界.

证明：记 $\boldsymbol{W} = T(X) - \boldsymbol{g}(\theta) - \boldsymbol{\Delta}\boldsymbol{I}^{-1}(\theta)\boldsymbol{S}_\theta(X)$，其期望与方差分别为

$$\mathrm{E}\boldsymbol{W} = \mathrm{E}_\theta(\boldsymbol{T}(X)) - \boldsymbol{g}(\theta) - \boldsymbol{\Delta}\boldsymbol{I}^{-1}(\theta)\mathrm{E}\boldsymbol{S}_\theta(X) = 0$$

$$\begin{aligned}\mathrm{Var}_\theta(\boldsymbol{W}) &= \mathrm{E}(\boldsymbol{W}\boldsymbol{W}') \\ &= \mathrm{E}[(\boldsymbol{T}(X) - \boldsymbol{g}(\theta))(\boldsymbol{T}(X) - \boldsymbol{g}(\theta))'] + \boldsymbol{\Delta}\boldsymbol{I}^{-1}\mathrm{E}(\boldsymbol{S}_\theta\boldsymbol{S}_\theta')\boldsymbol{I}^{-1}\boldsymbol{\Delta}' \\ &\quad - \mathrm{E}[(\boldsymbol{T}(X) - \boldsymbol{g}(\theta))\boldsymbol{S}_\theta']\boldsymbol{I}^{-1}\boldsymbol{\Delta}' - \boldsymbol{\Delta}\boldsymbol{I}^{-1}\mathrm{E}[\boldsymbol{S}_\theta(\boldsymbol{T}(X) - \boldsymbol{g}(\theta))']\end{aligned}$$

$$= \mathrm{Var}_\theta(\boldsymbol{T}(X)) + \boldsymbol{\Delta I}^{-1}\boldsymbol{\Delta}' - \boldsymbol{\Delta I}^{-1}\boldsymbol{\Delta}' - \boldsymbol{\Delta I}^{-1}\boldsymbol{\Delta}'$$
$$= \mathrm{Var}_\theta(\boldsymbol{T}(X)) - \boldsymbol{\Delta I}^{-1}\boldsymbol{\Delta}'$$

由于 $\mathrm{Var}_\theta(\boldsymbol{W})\geqslant 0$，立有(2.13)式成立.

$s=k=1$ 是最简单而常见的情形，此时(2.13)式简化为

$$\mathrm{Var}_\theta(T(X)) \geqslant \left(\frac{\partial g(\theta)}{\partial\theta}\right)^2 / I(\theta) \tag{2.14}$$

对重复抽样结构，$I(\theta)=nI_1(\theta)$.

对信息不等式，我们必须牢记它是有条件的. 当定理条件不满足时，有这样的无偏估计存在，其方差比直接算得的 C-R 下界还要小. Cramer 于 1946 年给出如下的说明例子.

例 2.18 设 $X_1,\cdots,X_n$ 是来自密度函数为 $p_\theta(x)=\mathrm{e}^{-(x-\theta)},x>\theta$ 的一个样本. 显然，该分布族支撑与 θ 有关，不是 Cramer-Rao 正则族. 如直接计算 C-R 下界，有

$$I_1(\theta) = \mathrm{E}\left(\frac{\partial \ln p_\theta(X)}{\partial\theta}\right)^2 = 1$$

从而 θ 的 C-R 下界为 $1/I(\theta)=1/n$. 另外，我们知道此处 $X_{(1)}$ 是充分统计量，取 $\hat\theta(X)=X_{(1)}-1/n$，则 $\hat\theta(X)$ 的概率密度函数为

$$p(t;\theta) = n\mathrm{e}^{-(nt-n\theta+1)} I_{(t+1/n>\theta)}$$

不难算得

$$\begin{aligned}
\mathrm{E}(\hat\theta(X)) &= \int_{\theta-1/n}^{\infty} nt\mathrm{e}^{-(nt-n\theta+1)}\mathrm{d}t \\
&= \int_{\theta-1/n}^{\infty}(nt-n\theta+1)\mathrm{e}^{-(nt-n\theta+1)}\mathrm{d}t + \theta - 1/n \\
&= \int_0^{\infty} ny\mathrm{e}^{-ny}\mathrm{d}y + \theta - 1/n = \theta \\
\mathrm{E}(\hat\theta(X))^2 &= \int_0^{\infty} n(y+\theta-1/n)^2\mathrm{e}^{-ny}\mathrm{d}y \\
&= 1/n^2 + \theta^2
\end{aligned}$$

故

$$\mathrm{Var}(\hat\theta(X)) = 1/n^2$$

由此，$\hat\theta(X)$ 是 θ 的无偏估计，而其方差小于 $1/n$. 之所以这样，原因是该

分布族 Fisher 信息不存在，而 $I(\theta)$ 只是形式计算的结果. 顺便指出，指数族是 Cramer-Rao 正则族，它满足定理 2.6 的条件. 因而对指数族，信息不等式是成立的.

既然(2.13)式或(2.14)式给出了无偏估计的方差下界，人们自然要问，这个下界能否达到？如果某个无偏估计的方差达到了 C-R 下界，那末它显然就是 UMVUE 了.

例 2.19　设 $X_1,\cdots,X_n$ 是来自 $b(1,\theta)$ 的一个样本，因二点分布是正则族，其 Fisher 信息存在. 由(2.11)式，

$$S_\theta(X_1)=\frac{X_1}{\theta}-\frac{1-X_1}{1-\theta}=\frac{X_1-\theta}{\theta(1-\theta)}$$

$$I_1(\theta)=\mathrm{E}\left(\frac{X_1-\theta}{\theta(1-\theta)}\right)^2=\frac{1}{\theta(1-\theta)}$$

如果我们要估计 θ，则由(2.14)式知其无偏估计的 C-R 下界为 $\theta(1-\theta)/n$. 现取 $\overline{X}$ 作为 θ 的估计，则显有 $\mathrm{Var}_\theta(\overline{X})=\theta(1-\theta)/n$. 故 $\overline{X}$ 的方差达到 θ 的 C-R 下界，这也表明 $\overline{X}$ 是 θ 的 UMVUE.

如果我们要估计 $g(\theta)=\theta^2$，则 $g(\theta)$ 的无偏估计的 C-R 下界为 $4\theta^3(1-\theta)/n$. 由例 2.9 知 $g(\theta)$ 的 UMVUE 是 $T(X)=\sum X_i(\sum X_i-1)/n(n-1)$. 记 $Y=\sum X_i$，则 $Y\sim b(n,\theta)$，$T(X)=Y(Y-1)/(n(n-1))$. 对小于 n 的 k，

$$\begin{aligned}
&\mathrm{E}[Y(Y-1)\cdots(Y-k)]\\
&=\sum_{i=0}^{n}i(i-1)\cdots(i-k)\binom{n}{i}\theta^i(1-\theta)^{n-i}\\
&=\sum_{i=k+1}^{n}\frac{n!}{(i-k-1)!(n-i)!}\theta^i(1-\theta)^{n-i}\\
&=n(n-1)\cdots(n-k)\theta^{k+1}\cdot\\
&\quad\sum_{j=0}^{n-k-1}\frac{(n-k-1)!}{j!(n-k-1-j)!}\theta^j(1-\theta)^{n-k-1-j}\\
&=n(n-1)\cdots(n-k)\theta^{k+1}
\end{aligned}$$

易计算

$$E[Y^2(Y-1)^2] = E\{Y(Y-1)[(Y-2)(Y-3)+4(Y-2)+2]\}$$
$$= n(n-1)[(n-2)(n-3)\theta^4+4(n-2)\theta^3+2\theta^2]$$

$$\mathrm{Var}(T(X)) = \frac{1}{n(n-1)}[(n-2)(n-3)\theta^4+4(n-2)\theta^3+2\theta^2]-\theta^4$$
$$= \frac{1}{n(n-1)}[4(n-2)\theta^3+2\theta^2-(4n-6)\theta^4]$$
$$= 4\theta^3(1-\theta)/n+2\theta^2(1-\theta)^2/n(n-1)$$
$$> 4\theta^3(1-\theta)/n$$

此估计的方差严格大于 θ^2 的无偏估计的 C-R 下界. 因 $T(X)$是 θ^2 的 UMVUE,知 θ^2 的所有无偏估计的方差均严格大于其 C-R 下界.

例 2.20 设 $X_1,\cdots,X_n$ 是来自 $N(0,\theta)$的一个样本,$\theta>0$. 由(2.12)式,$I_1(\theta)=-E_\theta\left\{\frac{1}{2\theta^2}-\frac{X_1^2}{\theta^3}\right\}=1/2\theta^2$,从而 $I(\theta)=n/2\theta^2$. 现设要估计 θ. 由(2.14),对 θ 的任一无偏估计 $\hat{\theta}(X)$,

$$\mathrm{Var}_\theta(\hat{\theta}(X)) \geqslant 2\theta^2/n$$

取 $\hat{\theta}(X)=\frac{1}{n}\sum X_i^2$,简单验算即知 $E(\hat{\theta})=\theta$, $\mathrm{Var}_\theta(\hat{\theta})=2\theta^2/n$,从而 $\hat{\theta}(X)$的方差达到 C-R 下界,它是 θ 的 UMVUE.

若要估计 $\sigma \triangleq g(\theta)=\sqrt{\theta}$,因 $\partial g(\theta)/\partial\theta=\frac{1}{2}\theta^{-1/2}$. 由(2.14),对 σ 的任一无偏估计 $T(X)$,有

$$\mathrm{Var}_\theta(T(X)) \geqslant \theta/2n \tag{2.15}$$

取 $T(X)=c_n(\sum X_i^2)^{1/2}$,因 $\sum X_i^2/\sigma^2 \sim \chi_n^2$,可算得 $E(\sqrt{\sum X_i^2}/\sigma) = \sqrt{2}\,\Gamma\left(\frac{n+1}{2}\right)\Big/\Gamma\left(\frac{n}{2}\right)$. 因此,选 $c_n=\Gamma\left(\frac{n}{2}\right)\Big/\left[\sqrt{2}\,\Gamma\left(\frac{n+1}{2}\right)\right]$可使 $T(X)$成为 σ的 UMVUE. 此时

$$\mathrm{Var}_\theta(T(X)) = c_n^2E(\sum X_i^2)-\theta=(nc_n^2-1)\theta \tag{2.16}$$

对一些 n,(2.16)与 σ 的无偏估计的 C-R 下界的比较列于表 2.1 中.

表 2.1 (2.16)与 C-R 下界的比较

n	1	2	5	10
nc_n^2-1	0.5706	0.2732	0.1405	0.0512
$[2n(nc_n^2-1)]^{-1}$	0.8760	0.9149	0.9572	0.9769

由此可见，$T(X)$的方差(2.16)式达不到其 C-R 下界，但随着 n 的增大，二者的差距越来越小. 用 Stirling 公式可证明，$2n(nc_n^2-1)\to 1(n\to\infty)$(见例 2.5).

本例中，$\sigma=\sqrt{\theta}$ 的无偏估计的 C-R 下界也可直接由 $I(\sigma)$给出. 事实上，由于

$$S_\sigma(X_1)=\frac{\partial \ln P_{\sigma^2}(X_1)}{\partial\sigma}=-\frac{1}{\sigma}+\frac{X_1^2}{\sigma^3}$$

所以

$$I_1(\sigma)=-\mathrm{E}_\sigma\left(\frac{\partial S_\sigma(X_1)}{\partial\sigma}\right)=-\frac{1}{\sigma^2}+\mathrm{E}_\sigma\frac{3X_1^2}{\sigma^4}=\frac{2}{\sigma^2}$$

由此也可得出(2.15)式.

§2.3.4 有效无偏估计

综上可见，对正则条件下的分布族，有些参数无偏估计的 C-R 下界可以达到，而有些则不能. 通常把达到 C-R 下界的无偏估计称为有效无偏估计，而把无偏估计的方差与其 C-R 下界之比的倒数称为该估计的效，如表 2.1 中最后一行.

定义 2.11 设$\{p_\theta(x),\theta\in\Theta\}$是 Cramer-Rao 正则族，$g(\theta)$是可估参数，$T(X)$是 $g(\theta)$的一个无偏估计，则称

$$(g'(\theta))^2I^{-1}(\theta)/\mathrm{Var}_\theta(T(X))$$

为估计 $T(X)$的效. 如果效等于 1，则称 $T(X)$为 $g(\theta)$的有效无偏估计.

下面讨论有效无偏估计存在的条件.

定理 2.7 设$\{p_\theta(x),\theta\in\Theta\}$为 Cramer-Rao 正则族，其 Fisher 信息 $I(\theta)$存在不为 0，又设 $g(\theta)$的一阶导数存在不为 0. 倘存在 $g(\theta)$的

无偏估计 $T(X)$，使得

$$\mathrm{Var}_\theta(T(X)) = (g'(\theta))^2 I^{-1}(\theta)$$

则存在实函数 $c(\theta)$，$d(x)$ 和 $h(\theta)$，使

$$\ln p_\theta(x) = h(\theta)T(x) + c(\theta) + d(x) \tag{2.17}$$

或

$$p_\theta(x) = \exp\{h(\theta)T(x) + c(\theta) + d(x)\}$$

证明：令 $W = T(x) - g(\theta) - g'(\theta)I^{-1}S_\theta(x)$，则 $\mathrm{E}W = 0$，$\mathrm{Var}_\theta(W) = \mathrm{Var}_\theta(T(X)) - (g'(\theta))^2 I^{-1} = 0$，从而 $W = 0$，a.s.．由于 I 与 $g'(\theta)$ 均不为 0，故

$$S_\theta(x) = \frac{I(\theta)}{g'(\theta)}(T(x) - g(\theta)) \triangleq m(\theta)T(x) - n(\theta) \tag{2.18}$$

其中，$m(\theta) = I(\theta)/g'(\theta)$，$n(\theta) = I(\theta)g(\theta)/g'(\theta)$．取定 θ_0，

$$\int_{\theta_0}^{\theta} S_\omega(x)\mathrm{d}\omega = \int_{\theta_0}^{\theta}[T(x)m(\omega) - n(\omega)]\mathrm{d}\omega$$

$$\triangleq h(\theta)T(x) + c(\theta)$$

记 $d(x) = \ln p_{\theta_0}(x)$，便有(2.17)式，定理得证．

定理 2.7 表明，有效无偏估计只在指数型分布族场合才可谈及，然而并非指数族中任一可估参数都有有效无偏估计，如例 2.19 和例2.20 中看到的那样，有些参数有有效无偏估计，有些则无．下面的定理讨论有效无偏估计存在的充要条件．

定理 2.8 设 $X = (X_1, \cdots, X_n)$ 的分布为指数型分布

$$p(x;\theta) = \exp\{\theta T(x) + c(\theta) + d(x)\}$$

则 $g(\theta)$ 存在有效无偏估计的充要条件是存在常数 a，b，使得 $g(\theta) = a\mathrm{E}_\theta T(X) + b$．

证明：先验证充分性．因

$$I(\theta) = -\mathrm{E}_\theta\left\{\frac{\partial^2 \ln p(x;\theta)}{\partial\theta^2}\right\} = -\frac{\partial^2 c(\theta)}{\partial\theta^2}$$

而 $g(\theta) = a\mathrm{E}T(X) + b = -a\dfrac{\partial c(\theta)}{\partial\theta} + b$，$\partial g(\theta)/\partial\theta = -a\dfrac{\partial^2 c(\theta)}{\partial\theta^2}$，所以 $g(\theta)$ 的无偏估计的 C-R 下界为 $-a^2\dfrac{\partial^2 c(\theta)}{\partial\theta^2}$．取 $g(\theta)$ 的估计为 $\hat{g}(x) =$

$aT(x)+b$,即知 $\hat{g}(x)$ 是 $g(\theta)$ 的无偏估计,且 $\mathrm{Var}_\theta(\hat{g}(X))=a^2\mathrm{Var}_\theta(T(X))=-a^2\dfrac{\partial^2 c(\theta)}{\partial\theta^2}$,达到 C-R 下界.

再证必要性.不妨设 $g(\theta)$ 不恒为常数,$\hat{g}(x)$ 为 $g(\theta)$ 的有效无偏估计,由(2.18)式知,存在这样的 $m(\theta)\neq 0$,使 $S_\theta(x)=m(\theta)(\hat{g}(x)-g(\theta))$,又 $S_\theta(x)=\dfrac{\partial c(\theta)}{\partial\theta}+T(x)$,故

$$m(\theta)(\hat{g}(x)-g(\theta))=\partial c(\theta)/\partial\theta+T(x),\quad \forall\,\theta\in\Theta$$

此即

$$\hat{g}(x)=\alpha(\theta)T(x)+\beta(\theta),\quad \forall\,\theta\in\Theta$$

因 $\hat{g}(x)$ 与 θ 无关,立知 $\alpha(\theta),\beta(\theta)$ 应与 θ 无关,记为 a,b. 即证得必要性.

对多参数情形有类似的结论[4],此处从略.

效的概念在不同场合有不同定义,最常见的效是 2.5 节中定义的.此处定义的效不常用,因为我们已经看到,大多数场合是不存在有效无偏估计的,亦即效不可能达到 1,这使得人们对如此定义的效产生疑问.因此,有人建议用 UMVUE 的方差取代 C-R 下界来定义无偏估计的效.

信息不等式在正则场合确实提供了一种证明 UMVUE 的方法,然而我们已经看到,这种证明方法只在指数型分布族下对少数参数适用,而且它不如上节给出的方法来得简便.因此可以说,信息不等式在证明某个无偏估计为 UMVUE 方面作用甚小.然而,信息不等式在大样本场合给出了一个很好的方差下界,这个下界是可以逼近的.由此引出了一类很重要的估计:渐近有效的估计序列.这将在 2.5 节加以讨论.

§2.4 矩估计与替换方法

§2.4.1 矩估计

前面三节主要从估计性质角度讨论了参数估计问题.本节和下几

节则讨论如何找出参数的估计方法.矩法,极大似然方法和最小二乘法是寻找参数点估计的三种主要方法.本节先介绍矩法.

用矩法求估计被认为是最古老的求估计的方法之一,它由K. Pearson在本世纪初提出.其基本思路是用样本矩及其函数估计相应的总体矩及其函数,具体如下.

设$X_1,\cdots,X_n$是来自某总体的一个样本,以μ_r记总体的r阶原点矩,m_r记由$X_1,\cdots,X_n$得到的r阶样本原点矩,即

$$\mu_r = EX_1^r; \quad m_r = \frac{1}{n}\sum X_i^r \tag{2.19}$$

如果某参数θ可以表示为总体前k阶矩的函数,即

$$\theta = g(\mu_1,\cdots,\mu_k)$$

则我们可用$\hat{\theta}(X)=g(m_1,\cdots,m_k)$估计$\theta$,$\hat{\theta}(X)$即称为$\theta$的矩估计.比如,由矩法,我们可用样本均值和样本方差分别估计总体均值和总体方差.

例2.21 设$X_1,\cdots,X_n$是来自$b(1,\theta)$的一个样本.如果θ表示某事件的成功概率,通常事件的成败机会比$g(\theta)=\theta/(1-\theta)$是人们感兴趣的参数.单纯从无偏性等角度导出$g(\theta)$的估计是困难的,事实上,此处$g(\theta)$没有无偏估计(见例2.7).然而我们可以由矩法轻松地给出$g(\theta)$的一个很不错的估计.因θ是总体均值,由矩法,记$\overline{X}=\frac{1}{n}\sum X_i$,则

$$T(X) = \overline{X}/(1-\overline{X})$$

是$g(\theta)$的一个矩估计.

另外,由于$\mathrm{Var}_\theta(X_1)=\theta(1-\theta)$,$g(\theta)$也可写成

$$g(\theta) = \theta^2/[\theta(1-\theta)]$$

从而$\overline{X}^2/S_n^2$也是$g(\theta)$的一个矩估计.

这种矩估计不唯一的现象不是个别的.比如,若$X_1,\cdots,X_n$是来自Poisson分布$P(\lambda)$的一个样本,由于$EX_1=\lambda$,$\mathrm{Var}(X_1)=\lambda$,因此,样本均值$\overline{X}$和样本方差$S_n^2$都可以作为$\lambda$的矩估计;又比如,对正态分布$N(\mu,\sigma^2)$,由于其偶数阶中心矩为

$$\mu'_{2r} = \frac{\sigma^{2r}}{2^r}\frac{(2r)!}{r!}$$

因此任意一个 $2r$ 阶样本中心矩都可以给出一个 σ^2 的矩估计.

在矩估计不唯一时，人们可以根据如下两个基本原则来选择矩估计：(1)涉及到的矩的阶数尽可能小，从而对总体的要求也尽可能少；常用的矩估计一般只涉及一、二阶矩. (2)所用估计最好是(最小)充分统计量的函数，因为充分性原则在各种统计问题中都应是适合的. 譬如，在例 2.21 中，对二点分布，$\theta/(1-\theta)$ 的矩估计应取 $\overline{X}/(1-\overline{X})$；对 Poisson 分布 $P(\lambda)$，λ 的矩估计应取 $\overline{X}$；而正态分布中 σ^2 的矩估计自然应取二阶样本中心矩(样本方差).

§2.4.2　矩估计的特点

矩估计有两个基本特点，其一，矩估计基于经验分布函数，而经验分布函数逼近真实分布的前提条件是样本容量较大，因而理论上讲，矩方法是以大样本为应用对象的；其二，矩方法没有用到总体分布的任何信息，本质上讲它是一种非参数方法，对已知的总体分布，矩估计不一定是一个好的估计. 虽然如此，人们一般并不把矩方法完全当成非参数方法，它在小样本场合也常得到应用. 事实上，在许多场合，比如对二项分布，Poisson 分布，正态分布等常见分布的均值，其矩估计具有许多优良的性质，它是 UMVUE.

对一般情况下的矩估计，通常只能考虑它的大样本性质. 下面的定理给出了矩估计的大样本性质.

定理 2.9　设 $X_1,\cdots,X_n$ 为独立同分布变量序列，$\mathrm{E}|X_1|^k<\infty$，$\boldsymbol{\theta}=\boldsymbol{g}(\mu_1,\cdots,\mu_k)$，并设 $\boldsymbol{g}$ 是连续的，则矩估计 $\hat{\boldsymbol{\theta}}=\boldsymbol{g}(m_1,\cdots,m_k)$ 是 $\boldsymbol{\theta}$ 的相合估计.

证明：由大数定律、定理 2.1 及 $\boldsymbol{g}$ 的连续性可得.

定理 2.10　设 $X_1,\cdots,X_n$ 为独立同分布变量序列，$\mathrm{E}|X_1|^{2k}<\infty$. 记 $\boldsymbol{\theta}=(\theta_1,\cdots,\theta_s)'$，$\theta_i=g_i(\mu_1,\cdots,\mu_k)$，并设 g_i 对 μ_j 有连续偏导数，$i=1,\cdots,s$，$j=1,\cdots,k$. 记 $\boldsymbol{G}=(\partial g_i/\partial\mu_j)_{s\times k}$，则对 $\boldsymbol{\theta}$ 的矩估计 $\hat{\boldsymbol{\theta}}=(g_1(m_1,\cdots,m_k),\cdots,g_s(m_1,\cdots,m_k))'$，有

$$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) \xrightarrow{L} N_s(0, \boldsymbol{G\Sigma G'})$$

其中 Σ 是 $k \times k$ 阶矩阵，其(i,j)元素为 $\mu_{i+j}-\mu_i\mu_j$.

证明：这是下面引理 2.12 的特例.

引理 2.11 记 $\boldsymbol{T}_n=(T_{1n},\cdots,T_{kn})'$，$\boldsymbol{\theta}=(\theta_1,\cdots,\theta_k)'$，设

$$\sqrt{n}(\boldsymbol{T}_n-\boldsymbol{\theta}) \xrightarrow{L} N(0,\boldsymbol{\Sigma})$$

其中 $\boldsymbol{\Sigma}=(\sigma_{ij})_{k\times k}$. 又设 $\boldsymbol{g}(t_1,\cdots,t_k)$对各 t_i 有连续偏导数，则当 $n\to\infty$时有

$$\sqrt{n}[\boldsymbol{g}(T_{1n},\cdots,T_{kn})-\boldsymbol{g}(\theta_1,\cdots,\theta_k)] \xrightarrow{L} N(0,\sigma^2(\boldsymbol{\theta}))$$

其中 $\sigma^2(\theta)=\sum\sum\left(\frac{\partial g}{\partial\theta_i}\frac{\partial g}{\partial\theta_j}\sigma_{ij}\right)$.

证明：由 Taylor 展开式

$$\boldsymbol{g}(T_{1n},\cdots,T_{kn})-\boldsymbol{g}(\theta_1,\cdots,\theta_k)=\sum_{i=1}^{k}(T_{in}-\theta_i)\frac{\partial g}{\partial\theta_i}+\varepsilon_n\|\boldsymbol{T}_n-\boldsymbol{\theta}\|$$

其中 $\|\boldsymbol{T}_n-\boldsymbol{\theta}\|=\sqrt{(\boldsymbol{T}_n-\boldsymbol{\theta})'(\boldsymbol{T}_n-\boldsymbol{\theta})}$，且当 $\boldsymbol{T}_n\to\theta$ 时 $\varepsilon_n\to 0$，亦即，$\forall\ \varepsilon>0$，$\exists\ \delta>0$，只要 $|T_{in}-\theta_i|<\delta, i=1,\cdots,k$，有$|\varepsilon_n|<\varepsilon$，于是

$$Pr(|\varepsilon_n|<\varepsilon)\geqslant Pr(|T_{in}-\theta_i|<\delta, i=1,\cdots,k)\to 1$$

即 $\varepsilon_n \xrightarrow{P} 0$，又 $\sqrt{n}(\boldsymbol{T}_n-\boldsymbol{\theta})$依分布收敛，立知

$$\sqrt{n}\varepsilon_n\|\boldsymbol{T}_n-\boldsymbol{\theta}\| \xrightarrow{P} 0$$

由 Slutskey 定理，$\sqrt{n}[\boldsymbol{g}(T_{1n},\cdots,T_{kn})-\boldsymbol{g}(\theta_1,\cdots,\theta_k)]$与 $\sqrt{n}\sum_{i=1}^{k}\frac{\partial g}{\partial\theta_i}(T_{in}-\theta_i)$有相同的渐近分布，其方差为 $\sigma^2(\theta)=\sum\sum\left(\frac{\partial g}{\partial\theta_i}\frac{\partial g}{\partial\theta_j}\sigma_{ij}\right)$，引理得证.

引理 2.12 $\boldsymbol{T}_n$ 及其渐近分布同引理 2.11，设向量函数

$$\boldsymbol{g}(t_1,\cdots,t_k)=(g_1(t_1,\cdots,t_k),\cdots,g_s(t_1,\cdots,t_k))'$$

其中每一个 g_i 对 t_j 都有连续偏导数，记 $\boldsymbol{G}=\left(\frac{\partial g_i}{\partial\theta_j}\right)_{s\times k}$，则有

$$\sqrt{n}[\boldsymbol{g}(T_{1n},\cdots,T_{kn})-\boldsymbol{g}(\theta_1,\cdots,\theta_k)] \xrightarrow{L} N_s(0,\boldsymbol{G\Sigma G'}) \quad (2.20)$$

证明：对任意的 $\boldsymbol{a}=(a_1,\cdots,a_s)'\in\boldsymbol{R}^s$，令 $\delta=\sum_{i=1}^{s}a_ig_i$，则由引理

2.11 知

$$\sqrt{n}\left[\delta(T_{1n},\cdots,T_{kn})-\delta(\theta_1,\cdots,\theta_k)\right]\xrightarrow{L}N(0,\sigma_\delta^2)$$

其中

$$\begin{aligned}\sigma_\delta^2&=\sum_i\sum_j\left\{\frac{\partial\delta}{\partial\theta_i}\frac{\partial\delta}{\partial\theta_j}\sigma_{ij}\right\}\\&=\sum_i\sum_j\left\{\left(\sum_r a_r\frac{\partial g_r}{\partial\theta_i}\right)\left(\sum_l a_l\frac{\partial g_l}{\partial\theta_j}\right)\sigma_{ij}\right\}\\&=\boldsymbol{a}'\boldsymbol{G\Sigma G'}\boldsymbol{a}\end{aligned}$$

由 $\boldsymbol{a}$ 的任意性知(2.20)式成立.

有了定理 2.10,我们可以在相当一般的条件下求矩估计的渐近分布.

例 2.22　设 $X_1,\cdots,X_n$ 为独立同分布变量,$\mathrm{E}|X_1|^4<\infty$,记 $\boldsymbol{\theta}=(\mu,\sigma^2)'$,其中

$$\mu=\mathrm{E}X_1=\mu_1\triangleq g_1(\mu_1,\mu_2)$$
$$\sigma^2=\mathrm{Var}(X_1)=\mu_2-\mu_1^2\triangleq g_2(\mu_1,\mu_2)$$

于是

$$\boldsymbol{G}=\left(\frac{\partial g_i}{\partial\mu_j}\right)_{2\times2}=\begin{pmatrix}1&0\\-2\mu_1&1\end{pmatrix}=\begin{pmatrix}1&0\\-2\mu&1\end{pmatrix}$$

令 $\boldsymbol{T}_n=(g_1(m_1,m_2),g_2(m_1,m_2))'=(\overline{X}_n,S_n^2)'$,由定理 2.10 有

$$\sqrt{n}(\boldsymbol{T}_n-\boldsymbol{\theta})\xrightarrow{L}N_2(0,\boldsymbol{G\Sigma G'})$$

其中

$$\boldsymbol{\Sigma}=\begin{pmatrix}\sigma^2&\mu_3-\mu_1\mu_2\\\mu_3-\mu_1\mu_2&\mu_4-\mu_2^2\end{pmatrix}$$

经计算得

$$\boldsymbol{G\Sigma G'}=\begin{pmatrix}\sigma^2&\nu_3\\\nu_3&\nu_4-\sigma^4\end{pmatrix}$$

其中 ν_j 表示 X_1 的 j 阶中心矩,即 $\nu_j=\mathrm{E}(X_1-\mu)^j$,$j=3,4$. 由此我们可以得到 $\overline{X}$,S_n^2 及其函数的渐近分布. 譬如,由

$$\sqrt{n}(S_n^2-\sigma^2)\xrightarrow{L}N(0,\nu_4-\sigma^4)$$

令 $h(t)=\sqrt{t}$,则对 σ 的矩估计 S_n 有

$$\sqrt{n}(S_n - \sigma) \xrightarrow{L} N(0, (\nu_4 - \sigma^4)/4\sigma^2)$$

在讨论矩估计的大样本性质时，其相应的总体矩的存在性是很重要的. 若总体矩不存在，样本矩也不是相合的，即使是一阶样本矩，Cauchy 分布是众知的例.

例 2.23　设 $X_1, \cdots, X_n$ 是来自 Cauchy 分布的一个样本，密度函数为

$$p(x;\theta) = \{\pi[1 + (x-\theta)^2]\}^{-1} \tag{2.21}$$

因 Cauchy 分布的特征函数为 $e^{it\theta - |t|}$，易知 $\overline{X}_n = \frac{1}{n}\sum X_i$ 的特征函数也为 $e^{it\theta - |t|}$，故 $\overline{X}_n$ 的分布仍为(2.21)，从而，$\forall\ \delta > 0, Pr(|\overline{X}_n - \theta| > \delta)$ 对所有 n 保持不变，故 $\overline{X}_n$ 不是 θ 的相合估计，事实上，$\overline{X}_n$ 不是任何一个参数的相合估计.

§ 2.4.3　频率替换估计

考虑由 n 次独立重复试验组成的序列，每次试验有 k 种可能的试验结果出现. 令 p_j 表示出现第 j 个结果的概率，$\sum_{j=1}^k p_j = 1$. 设 n_j 表示 n 次试验中结果 j 出现的次数，则 $\boldsymbol{n} = (n_1, \cdots, n_k)'$ 称为单元频数向量，其精确分布为多项分布 $M(n, p_1, \cdots, p_k)$，显然，可用 n_j/n 估计 p_j. 而且，令 $\boldsymbol{X}_i = (0, \cdots, 0, 1, 0, \cdots, 0)'_{k\times 1}$，它表示第 i 次试验的结果. 若试验结果为 j，则 $\boldsymbol{X}_i$ 中唯一的一个 1 位于第 j 位，则 $\boldsymbol{n} = \sum_{i=1}^n \boldsymbol{X}_i$，因此，$\boldsymbol{n}$ 是 iid 和，由中心极限定理立有

$$\sqrt{n}\left(\frac{\boldsymbol{n}}{n} - \boldsymbol{p}\right) \xrightarrow{L} N_k(0, \boldsymbol{\Sigma}) \tag{2.22}$$

其中 $\boldsymbol{p} = (p_1, \cdots, p_k)'$，$\boldsymbol{\Sigma} = (\sigma_{ij})_{k\times k}$，

$$\sigma_{ij} = \begin{cases} p_j(1-p_j), & i = j \\ -p_i p_j, & i \neq j \end{cases}$$

进一步，$\forall\ \boldsymbol{b} = (b_1, \cdots, b_k) \in \boldsymbol{R}^k$，我们可用 $\boldsymbol{b}'\boldsymbol{n}$ 估计 $\boldsymbol{b}'\boldsymbol{p}$，且

$$\sqrt{n}(\boldsymbol{b}'\boldsymbol{n} - \boldsymbol{b}'\boldsymbol{p}) \xrightarrow{L} N(0, \sigma_b^2)$$

其中 $\sigma_b^2 = b'\Sigma b = \sum p_j b_j^2 - (\sum p_j b_j)^2$.

选取不同的 b,我们可以对一些感兴趣的参数进行估计. 比如估计具有某种性质的结果出现的概率,两种具不同性质的结果出现概率之差等等.

更一般地,我们可以考虑 $p_1,\cdots,p_k$ 某个连续函数的估计,由引理 2.11 和引理 2.12 知,只要该函数关于 p_j 的偏导数存在且连续,则该估计也是渐近正态的.

常见的情形是,p_j 是参数 $\theta=(\theta_1,\cdots,\theta_m)'$ 的连续函数. 如果 θ 能够表示为 $p_1,\cdots,p_k$ 的连续函数 $h(p_1,\cdots,p_k)$,则我们可用 $h\left(\frac{n_1}{n},\cdots,\frac{n_k}{n}\right)$ 估计 θ. 当然,θ 的这种表示可能不唯一,从而对 θ 有不同的估计,存在着一个估计的选择问题. 下面我们用一个例子来说明这一问题.

例 2.24 假设试验有三种可能结果,分别记为 1,2,3,其出现概率分别为

$$p_1 = \theta^2, p_2 = 2\theta(1-\theta), p_3 = (1-\theta)^2$$

其中 $0<\theta<1$,这是著名的 Hardy-Weinberg 模型.

记 n_i 为 n 次试验中结果 i 出现的次数,则 $(n_1,n_2,n_3)'$ 服从 $M(n,p_1,p_2,p_3)$. 现我们要估计 θ,我们可以把 θ 表示成 p_1,p_2,p_3 的连续函数. 此处这种表示不唯一,比如,$\theta=\sqrt{p_1}$,$\theta=1-\sqrt{p_3}$,$\theta=p_1+p_2/2$ 等等. 因而我们至少有三个看起来合理的估计

$$T_1 = \sqrt{n_1/n}, \quad T_2 = 1-\sqrt{n_3/n}, \quad T_3 = (n_1+n_2/2)/n$$

对 T_1,T_2,T_3 的比较当然可以根据均方误差进行,然而这里均方误差的计算相当繁杂很难完成,而渐近性则是简单而有效的比较准则. 显然,T_1,T_2,T_3 都是 θ 的相合估计,其渐近正态分布也是可求的,这时渐近方差大小就是一个选择标准.

对 $\theta=h(p_1,\cdots,p_k)$,由(2.22)式及引理 2.11 知

$$\sqrt{n}\left(h\left(\frac{n_1}{n},\cdots,\frac{n_k}{n}\right)-\theta\right) \xrightarrow{L} N(0,\sigma_h^2) \tag{2.23}$$

其中

$$\sigma_h^2 = \sum_{i=1}^k \sum_{j=1}^k \frac{\partial h}{\partial p_i} \frac{\partial h}{\partial p_j} \sigma_{ij} = \sum_{j=1}^k p_j \left(\frac{\partial h}{\partial p_j} \right)^2 - \left(\sum_{j=1}^k p_j \frac{\partial h}{\partial p_j} \right)^2$$

本例中，$h_1 = \sqrt{p_1}$，$h_2 = 1 - \sqrt{p_3}$，$h_3 = p_1 + p_2/2$. 其对应的 σ_h^2 分别为

$$\sigma_1^2 = p_1 \left[\frac{1}{2\sqrt{p_1}} \right]^2 - \left[\frac{p_1}{2\sqrt{p_1}} \right]^2$$

$$= \frac{1}{4}(1 - p_1) = \frac{1}{4}(1 - \theta^2)$$

$$\sigma_2^2 = p_3 \left[\frac{1}{2\sqrt{p_3}} \right]^2 - \left[\frac{p_3}{2\sqrt{p_3}} \right]^2$$

$$= \frac{1}{4}(1 - p_3) = \frac{1}{4}[1 - (1 - \theta)^2]$$

$$\sigma_3^2 = (p_1 + p_2/4) - (p_1 + p_2/2)^2$$

$$= \theta^2 + \theta(1 - \theta)/2 - \theta^2 = \theta(1 - \theta)/2$$

由(2.23)，$\sqrt{n}(T_i - \theta) \xrightarrow{L} N(0, \sigma_i^2)$，$i = 1, 2, 3$.

对 $\sigma_1^2, \sigma_2^2, \sigma_3^2$ 进行比较可知 σ_3^2 最小，而 σ_1^2 与 σ_2^2 互有优劣，因而可以认为 T_3 是一个好的估计(见图 2.1). 事实上，T_3 即为下一节要讨论的 θ 的极大似然估计.

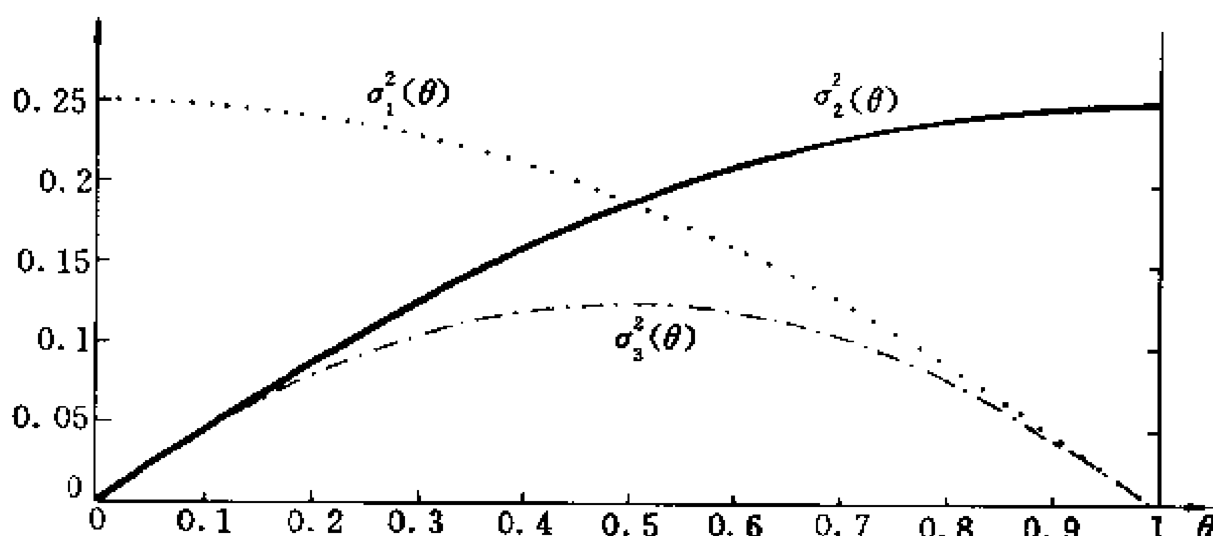


图 2.1 $\sigma_1^2, \sigma_2^2, \sigma_3^2$ 的比较

§ 2.5 极大似然估计

§ 2.5.1 定义与例子

极大似然方法是统计中最重要,应用最广泛的方法之一. 该方法最初由德国数学家 Gauss 于 1821 年提出,但未得到重视,R. A. Fisher 在 1922 年再次提出了极大似然的思想并探讨了它的性质,使之得到广泛研究和应用.

在概率统计中,概率密度函数 $p(x;\theta)$扮演了重要角色. 当 θ 已知时,$p(x;\theta)$显示概率密度怎样随 x 变化;反过来,当样本 x 给定后,可考虑对不同的 θ,概率密度如何变化,它反映了对 x 的解释能力,这便是似然.

定义 2.12 设 $p(x;\theta),\theta\in\Theta$ 是$(\mathbf{R}^n,\mathscr{B}_{\mathbf{R}^n})$上的一族联合概率密度函数,对给定的 x,称

$$L(\theta;x) = kp(x;\theta) \tag{2.24}$$

为 θ 的似然函数,其中 $k>0$ 是不依赖于 θ 的量,常取 $k=1$. 进一步,若存在$(\mathbf{R}^n,\mathscr{B}_{\mathbf{R}^n})$到$(\Theta,\mathscr{B}_\Theta)$的统计量 $\hat{\theta}(x)$使

$$L(\hat{\theta}(x);x) = \sup_\theta L(\theta;x) \tag{2.25}$$

则 $\hat{\theta}(x)$称为 θ 的一个极大似然估计(Maximum Likelihood Estimate),简称 MLE.

由于概率密度函数大多具有指数函数形式,采用似然函数的对数通常更为简便. 称

$$l(\theta;x) = \ln L(\theta;x) \tag{2.26}$$

为 θ 的对数似然函数. 由于对数变换是严格单调增的,故 $l(\theta;x)$与 $L(\theta;x)$在寻求极大值时是等价的.

当 MLE 存在时,寻找 MLE 最常用的方法是求导数. 如果 $\hat{\theta}(x)$是 Θ 的内点,则 $\hat{\theta}(x)$是下列似然方程

$$\partial l(\theta;x)/\partial\theta_i = 0, i = 1,\cdots,k \tag{2.27}$$

的解.

例 2.25　设 $X_1,\cdots,X_n$ 是来自 $N(\mu,\sigma^2)$的一个样本,则

$$l(\mu,\sigma^2;x) = -n\ln\sigma - \sum(x_i-\mu)^2/2\sigma^2$$

由(2.27)

$$\frac{\partial l}{\partial\mu} = \frac{1}{2\sigma^2}\left(\sum x_i - n\mu\right) = 0$$

$$\frac{\partial l}{\partial\sigma^2} = \frac{\sum(x_i-\mu)^2}{\sigma^4} - \frac{n}{2\sigma^2} = 0$$

由此给出 μ 和 σ^2 的估计为

$$\hat{\mu} = \bar{x},\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = s_n^2 \tag{2.28}$$

可直接验证,$\hat{\mu},\hat{\sigma}^2$ 确是 μ 和 σ^2 的 MLE.

如果我们对参数空间加以限制,譬如,假设已知 $\mu>0$,MLE 将发生变化.本例中,若 $\bar{x}>0$,则(2.28)仍是 μ 和 σ^2 的 MLE.若 $\bar{x}<0$,则由于 $\partial l/\partial\mu$ 恒小于 0,$l(\mu,\sigma^2;x)$关于 μ 单调降,易知当 μ 为 0 而 σ^2 为 $\sum x_i^2/n$ 时 $L(\mu,\sigma^2;x)$达到最大.因此 MLE 变为

$$\hat{\mu} = \max(\bar{x},0)$$

$$\hat{\sigma}^2 = \sum(x_i-\hat{\mu})^2/n$$

例 2.26　设 $X_1,\cdots,X_n$ 是来自 $b(1,\theta)$的一个样本,$0<\theta<1$

$$l(\theta;x) = \theta\ln\sum x_i + (1-\theta)\ln(n-\sum x_i)$$

由$\frac{\partial l}{\partial\theta}=0$ 知

$$\hat{\theta} = \sum x_i/n = \bar{x}$$

当 $0<\bar{x}<1$ 时,易验证 $\hat{\theta}=\bar{x}$ 是 θ 的唯一的 MLE,而当 $\sum x_i = 0$ 或 n 时,由于 $\bar{X}$ 不在 Θ 内,严格意义上讲 MLE 不存在,但在实用中,人们往往仍把 $\bar{X}$ 称为 θ 的 MLE.为此,有人对 MLE 定义加以补充:当 $\hat{\theta}(x)\notin\Theta$,若存在一串 $\hat{\theta}_n(x)$,使$\lim\limits_{n\to\infty}\hat{\theta}_n(x)=\hat{\theta}(x)$,且 $\hat{\theta}_n(x)\in\Theta,n=1,2,\cdots$,则$\hat{\theta}(x)$也称为 θ 的 MLE.

由上述两个例子还可以看到,MLE 常是充分统计量的函数.由因

子分解定理知，若 T 为充分统计量，则 $p(x;\theta)=g(\theta,T(x))h(x)$，故 $l(\theta;x)=\ln g(\theta,T(x))+\ln h(x)$. 由此，使 $l(\theta;x)$ 达最大与使 $g(\theta,T(x))$ 达最大是等价的，故 $\hat{\theta}(x)$ 可取为 $T(x)$ 的函数.

在上述两个例子中，似然方程的解唯一，这在指数型分布族中都成立. 简单起见，以单参数指数族为例说明. 设 $X_1,\cdots,X_n$ 为来自 $p(x;\theta)=\exp\{\theta T(x)+c(\theta)+d(x)\}$ 的一个样本，则

$$\frac{\partial l(\theta;x)}{\partial\theta}=\sum T(x_i)+n\frac{\partial c(\theta)}{\partial\theta}=0 \tag{2.29}$$

由于 $ET(X)=-\partial c(\theta)/\partial\theta$, $\mathrm{Var}_\theta T(X)=-\partial^2 c(\theta)/\partial\theta^2>0$，故(2.29)若有解，则此解唯一，且若 Θ 是自然参数空间，则该解即为 MLE. 如果(2.29)无解，则 MLE 在 Θ 的边界处达到. 对多参数指数族 $p(x;\theta)=\exp\{\sum_{j=1}^{k}\theta_j T_j(x)+c(\theta)+d(x)\}$，似然方程化为

$$\sum T_j(x_i)=-n\partial c(\theta)/\partial\theta_j \quad j=1,\cdots,k \tag{2.30}$$

有完全相同的结论.

MLE 有一个非常有吸引力的性质：如果 $\hat{\theta}$ 是 θ 的 MLE，$g(\cdot)$ 为可测函数，则 $g(\hat{\theta})$ 也是 $g(\theta)$ 的 MLE. 该性质称为 MLE 的不变性，它使 MLE 在用于参数函数时十分简便.

如果 $g(\cdot)$ 是一一对应的，不变性是显然的. 如果 $g(\cdot)$ 不是一一对应的，为衡量 $g(\theta)$ 的可能性，给出诱导似然的概念. 设 $g(\cdot)$ 是 Θ 到 Ω 的可测变换，对任意的 $\omega\in\Omega$，定义 $G(\omega)=\{\theta: g(\theta)=\omega\}$，$G(\omega)$ 称为 ω 的陪集，称 $M(\omega;x)=\sup\limits_{\theta\in G(\omega)}L(\theta;x)$ 为诱导似然函数，在此意义下，因 $\hat{\theta}\in G(g(\hat{\theta}))$，且

$$L(\hat{\theta};x)=\sup_{\theta\in\Theta}L(\theta;x)$$

故对任意的 $\omega\in\Omega$，有

$$\begin{aligned}M(g(\hat{\theta});x)&=\sup_{\theta\in G(g(\hat{\theta}))}L(\theta;x)\\&=L(\hat{\theta};x)=\sup_{\theta\in\Theta}L(\theta;x)\\&\geqslant\sup_{\theta\in G(\omega)}L(\theta;x)=M(\omega;x)\end{aligned}$$

这就证明了 MLE 具有不变性.

例 2.27 设$(X_1,Y_1),\cdots,(X_n,Y_n)$是来自二元正态总体的一个样本，$E(X_1)=E(Y_1)=0$，$\mathrm{Var}(X_1)=\mathrm{Var}(Y_1)=\sigma^2$，$\mathrm{Cov}(X_1,Y_1)=\rho\sigma^2$，$\sigma^2>0$，$\rho\in(-1,1)$，下求 σ^2,ρ 的 MLE. 因

$$p(x,y;\sigma^2,\rho)=\left[2\pi\sigma^2\sqrt{1-\rho^2}\right]^{-1}\cdot \exp\left\{-\frac{1}{2\sigma^2(1-\rho^2)}(x^2+y^2-2\rho xy)\right\}$$

引入 $\theta_1=(\sigma^2(1-\rho^2))^{-1}>0$，$\theta_2=\rho\theta_1$，则 $\rho=\theta_2/\theta_1$，$\left[\sigma^2\sqrt{1-\rho^2}\right]^{-1}=\sqrt{\theta_1^2-\theta_2^2}$. 于是

$$p(x,y;\theta_1,\theta_2)=\exp\left\{\theta_1\left[-\frac{1}{2}(x^2+y^2)\right]+\theta_2 xy+\frac{1}{2}\ln(\theta_1^2-\theta_2^2)-\ln(2\pi)\right\}$$

这是指数族，其 $c(\theta_1,\theta_2)=\dfrac{n}{2}\ln(\theta_1^2-\theta_2^2)$，由(2.30)可得似然方程

$$\frac{1}{2}\,\frac{2\theta_1}{\theta_1^2-\theta_2^2}=\frac{1}{2n}\left(\sum X_i^2+\sum Y_i^2\right)$$

$$\frac{1}{2}\,\frac{2\theta_2}{\theta_1^2-\theta_2^2}=\frac{1}{n}\left(\sum X_iY_i\right)$$

利用 MLE 的不变性，将 θ_1,θ_2 用 σ^2,ρ 表示，可以方便地得到 σ^2、ρ 的 MLE 为

$$\hat{\sigma}^2=\frac{1}{2n}\sum(x_i^2+y_i^2)$$

$$\hat{\rho}=2\sum x_iy_i\Big/\sum(x_i^2+y_i^2)$$

§2.5.2 相合性与渐近正态性

作为一种统计思想，似然方法有其合理性，得到人们认可. 然而就小样本而言，却没有一个定理能说明 MLE 本身有什么优良性质. MLE 的优良性主要体现在大样本场合，在大样本理论中，MLE 扮演了一个中心角色. 下面几个定理给出了 MLE 的大样本性质.

设 $X_1,\cdots,X_n$ 是来自 $p(x;\theta)$的一个样本. 简单起见，考虑单参数情形，设参数空间是一个开区间，$l(\theta;x)=\sum_{i=1}^n\ln p(x_i;\theta)$.

定理 2.13　若 $\ln p(x;\theta)$ 在 Θ 上可微，并设 $p(x;\theta)$ 是可识别的（$\forall\ \theta\neq\theta'$，$\{x:p(x;\theta)\neq p(x;\theta')\}$ 不是零测集），则似然方程在 $n\to\infty$ 时以概率 1 有解，且此解关于 θ 是相合的.

证明：$\forall\ \theta'\neq\theta$，因 $p(x;\theta)$ 是可识别的，由 Jensen 不等式

$$E_\theta\left(\ln\frac{p(x;\theta')}{p(x;\theta)}\right)<\ln\left[E_\theta\frac{p(x;\theta')}{p(x;\theta)}\right]=0$$

记参数真值为 θ_0，对充分小的 $\delta>0$，有 $(\theta_0-\delta,\theta_0+\delta)\subset\Theta$，且

$$E_{\theta_0}\ln\frac{p(x;\theta_0-\delta)}{p(x;\theta_0)}<0$$

$$E_{\theta_0}\ln\frac{p(x;\theta_0+\delta)}{p(x;\theta_0)}<0$$

由强大数定律，当 $n\to\infty$ 时，

$$\frac{1}{n}[l(\theta_0-\delta;x)-l(\theta_0;x)]\xrightarrow{\text{a.s.}}E_{\theta_0}\ln\frac{p(x;\theta_0-\delta)}{p(x;\theta_0)}<0$$

$$\frac{1}{n}[l(\theta_0+\delta;x)-l(\theta_0;x)]\xrightarrow{\text{a.s.}}E_{\theta_0}\ln\frac{p(x;\theta_0+\delta)}{p(x;\theta_0)}<0$$

又由于 $l(\theta;x)$ 在 $(\theta_0-\delta,\theta_0+\delta)$ 上连续，因而必有一局部最大点，记其为 $\hat{\theta}$，由于 $l(\theta;x)$ 可微，故 $\left.\frac{\partial l}{\partial\theta}\right|_{\hat{\theta}}=0$，从而，$n\to\infty$ 时似然方程以概率 1 有解，且 $|\hat{\theta}-\theta|<\delta$，由 δ 的任意性知 $\hat{\theta}$ 是关于 θ_0 相合的，证毕.

定理 2.14　假设 Θ 为开区间，概率密度函数 $p(x;\theta)$，$\theta\in\Theta$ 满足

(1)在参数真值 θ_0 的邻域内，$\partial\ln p/\partial\theta$，$\partial^2\ln p/\partial\theta^2$，$\partial^3\ln p/\partial\theta^3$ 对所有 x 都存在；

(2)在参数真值 θ_0 的邻域内，$|\partial^3\ln p/\partial\theta^3|\leqslant H(x)$，且 $EH(x)<\infty$；

(3)在参数真值 θ_0 处，

$$E_{\theta_0}\left[\frac{p'(X,\theta_0)}{p(X,\theta_0)}\right]=0\quad E_{\theta_0}\left[\frac{p''(X,\theta_0)}{p(X,\theta_0)}\right]=0$$

$$I(\theta_0)=E_{\theta_0}\left[\frac{p'(X,\theta_0)}{p(X,\theta_0)}\right]^2>0$$

其中撇号表示对 θ 的微分. 记 $\hat{\theta}_n$ 为 $n\to\infty$ 时似然方程的相合解，则

$$\sqrt{n}(\hat{\theta}_n-\theta_0)\xrightarrow{L}N(0,I^{-1}(\theta_0))$$

证明：由定理 2.13 知，$n\to\infty$时似然方程以概率 1 有解，记为 $\hat{\theta}_n$，$\hat{\theta}_n\xrightarrow{P}\theta_0$，其中 θ_0 为参数真值. 将$\dfrac{\partial l}{\partial\theta}$在 θ_0 处进行 Taylor 展开，有

$$\frac{\partial l}{\partial\theta}=\frac{\partial l}{\partial\theta}\bigg|_{\theta_0}+(\theta-\theta_0)\frac{\partial^2 l}{\partial\theta^2}\bigg|_{\theta_0}+\frac{(\theta-\theta_0)^2}{2}\frac{\partial^3 l}{\partial\theta^3}\bigg|_{\theta_1}$$

其中 θ_1 介于 θ 与 θ_0 之间. 取 $\theta=\hat{\theta}_n$，

$$0=\frac{\partial l}{\partial\theta}\bigg|_{\hat{\theta}_n}=\frac{\partial l}{\partial\theta}\bigg|_{\theta_0}+(\hat{\theta}_n-\theta_0)\frac{\partial^2 l}{\partial\theta^2}\bigg|_{\theta_0}+\frac{(\hat{\theta}_n-\theta_0)^2}{2}\frac{\partial^3 l}{\partial\theta^3}\bigg|_{\theta_1}$$

其中 θ_1 介于 θ_0 与 $\hat{\theta}_n$ 之间，故 $\theta_1\xrightarrow{P}\theta_0(n\to\infty)$. 于是，由上式得

$$\sqrt{n}(\hat{\theta}_n-\theta_0)=\frac{-\sqrt{n}\dfrac{1}{n}\dfrac{\partial l}{\partial\theta}\bigg|_{\theta_0}}{\dfrac{1}{n}\left\{\dfrac{\partial^2 l}{\partial\theta^2}\bigg|_{\theta_0}+\dfrac{(\hat{\theta}_n-\theta_0)}{2}\dfrac{\partial^3 l}{\partial\theta^3}\bigg|_{\theta_1}\right\}}\tag{2.31}$$

因 $l=\sum_{i=1}^n \ln p(x_i;\theta)$，故$\dfrac{\partial l}{\partial\theta}\bigg|_{\theta_0}$，$\dfrac{\partial^2 l}{\partial\theta^2}\bigg|_{\theta_0}$为独立同分布变量和. 由条件(3)有

$$E_{\theta_0}\left[\frac{\partial\ln p(X_1;\theta)}{\partial\theta}\bigg|_{\theta_0}\right]=0$$

$$\mathrm{Var}_{\theta_0}\left[\frac{\partial\ln p(X_1;\theta)}{\partial\theta}\bigg|_{\theta_0}\right]=I(\theta_0)$$

从而由中心极限定理有

$$\sqrt{n}\,\frac{1}{n}\frac{\partial l}{\partial\theta}\bigg|_{\theta_0}\xrightarrow{L}N(0,I(\theta_0))\tag{2.32}$$

又

$$E_{\theta_0}\left[\frac{\partial^2\ln p(X_1;\theta)}{\partial\theta^2}\bigg|_{\theta_0}\right]=E_{\theta_0}\left[\frac{p''(X_1,\theta_0)}{p(X_1,\theta_0)}-\left(\frac{p'(X_1,\theta_0)}{p(X_1,\theta_0)}\right)^2\right]$$
$$=-I(\theta_0)$$

由强大数定律知

$$\frac{1}{n}\frac{\partial^2 l}{\partial\theta^2}\bigg|_{\theta_0}\xrightarrow{a.s.}-I(\theta_0)\tag{2.33}$$

又当 $n\to\infty$ 时，θ_1 在 θ_0 的邻域内，$\left|\dfrac{\partial^3\ln p(x;\theta)}{\partial\theta^3}\bigg|_{\theta_1}\right|<H(x)$，于是

$$\frac{1}{n}\frac{\partial^3 l}{\partial\theta^3}\bigg|_{\theta_1}=O_p(1),$$

$$\frac{1}{n}\frac{(\hat{\theta}_n-\theta_0)}{2}\frac{\partial^3 l}{\partial\theta^3}\bigg|_{\theta_1}\xrightarrow{P}0 \tag{2.34}$$

由(2.33)和(2.34)知(2.31)式的分母依概率收敛到 $-I(\theta_0)$，而(2.32)表明(2.31)式的分子依分布收敛到 $N(0,I(\theta_0))$，应用 Slustkey 定理，定理得证.

定理 2.14 表明，在 Cramer-Rao 正则族场合，定理条件一般是满足的，从而 MLE 是渐近正态的. 当然，在非正则族场合，MLE 也可能是渐近正态的，正则条件并非必要条件，看下例.

例 2.28　设 $X_1,\cdots,X_n$ 为来自下述总体的一个样本，$X_1\sim\frac{1}{2}e^{-|x-\theta|}$，$\theta\in\boldsymbol{R}$，$x\in\boldsymbol{R}$，该分布族不是 Cramer-Rao 正则族，其对数似然函数为

$$l(\theta;x)=-n\ln 2-\sum_{i=1}^{n}|x_i-\theta|$$

记 $X_{(1)},\cdots,X_{(n)}$ 为样本的次序统计量，上述对数似然函数可分段写为

$$l(\theta;x)=\begin{cases}-n\ln 2-\sum_{i=1}^{n}x_i+n\theta, & \theta<x_{(1)}\\ \cdots\cdots & \\ -n\ln 2+\sum_{i=1}^{m}x_{(i)}-\sum_{i=m+1}^{n}x_{(i)} & x_{(m)}\leqslant\theta<x_{(m+1)}\\ +(n-2m)\theta, & \\ \cdots\cdots & \\ -n\ln 2-n\theta+\sum_{i=1}^{n}x_i, & \theta\geqslant x_{(n)}\end{cases}$$

显然，当 $2m>n$ 时 $l(\theta;x)$是 θ 的严格增函数；而当 $2m<n$ 时 $l(\theta;x)$是 θ 的严格减函数. 由此，如果 n 是奇数，$x_{\left(\frac{n+1}{2}\right)}$ 是 θ 的 MLE，如果 n 是偶数，则 $x_{\left(\frac{n}{2}\right)}$ 与 $x_{\left(\frac{n}{2}+1\right)}$ 之间任一值都是 θ 的 MLE，譬如可取样本中位数 $\frac{1}{2}\left\{x_{\left(\frac{n}{2}\right)}+x_{\left(\frac{n}{2}+1\right)}\right\}$. 由此，样本中位数 $m_{0.5}$ 是 θ 的 MLE，对此分布族，θ 也是总体中位数，由定理 1.8 知

$$\sqrt{n}(m_{0.5}-\theta)\xrightarrow{L}N(0,1)$$

通常，MLE 被认为是渐近正态的，但也有反例．比如，对$(0,\theta)$上的均匀分布，MLE 是$x_{(n)}$，而众所周知，$x_{(n)}$的渐近分布是极值分布，不是正态分布．MLE 甚至也可能不是相合估计，如下述反例．

例 2.29 设$X_1,\cdots,X_n$是来自二点分布的一个样本，其分布为

$$Pr(X_1=1)=1-Pr(X_1=0)=\begin{cases}\theta, & \theta\text{ 为有理数}\\ 1-\theta, & \theta\text{ 为无理数}\end{cases}$$

其中$0<\theta<1$．记$T_n=\sum x_i$，则似然函数

$$L(\theta;x)=\begin{cases}\theta^{T_n}(1-\theta)^{n-T_n}, & \theta\text{ 为有理数}\\ \theta^{n-T_n}(1-\theta)^{T_n}, & \theta\text{ 为无理数}\end{cases}$$

容易看出，θ的 MLE 为$\hat{\theta}=T_n/n$，即

$$L(\hat{\theta};x)=\sup_{\theta\in(0,1)}L(\theta;x)$$

但

$$\frac{T_n}{n}\xrightarrow{P}\begin{cases}\theta, & \theta\text{ 为有理数}\\ 1-\theta, & \theta\text{ 为无理数}\end{cases}\quad(n\to\infty)$$

由此，θ的 MLE T_n/n不是θ的相合估计．

关于定理 2.14 还有一点要指出，$I^{-1}(\theta)$是$\sqrt{n}\,(\hat{\theta}-\theta)$的渐近分布的方差，一般而言，它并不一定是$\sqrt{n}\,\hat{\theta}$的精确方差的极限．广言之，若$T_n$依分布收敛到$X$，并不一定有$\mathrm{Var}(T_n)\to\mathrm{Var}(X)$．

例 2.30 设$X\sim U(-1,1)$，令

$$T_n=\begin{cases}X, & |X|\leqslant 1-\dfrac{1}{n}\\ n, & |X|>1-\dfrac{1}{n}\end{cases}$$

则对任意的$x\in(-1,1)$，

$$F_{T_n}(x)=P(T_n\leqslant x,|X|\leqslant 1-1/n)+P(T_n\leqslant x,|X|>1-1/n)$$

其中

$$P(T_n\leqslant x,|X|>1-/n)\leqslant P(|X|>1-1/n)=1/n\to 0$$

而

$$P(T_n \leqslant x, |X| \leqslant 1 - 1/n) = \begin{cases} 0, & x \leqslant -(1-1/n) \\ \dfrac{x+1}{2} - \dfrac{1}{2n}, & |x| < 1 - 1/n \\ 1 - 1/n, & x \geqslant 1 - 1/n \end{cases}$$

$$\xrightarrow{n\to\infty} \begin{cases} 0, & x \leqslant -1 \\ \dfrac{x+1}{2}, & -1 < x < 1 \\ 1, & x \geqslant 1 \end{cases}$$

最后的函数不是别的，正是 X 的分布函数，所以 $T_n \xrightarrow{L} X$；另一方面，我们考察 T_n 的一、二阶矩，

$$\mathrm{E}T_n = n \cdot \frac{1}{n} + \int_{-1-1/n}^{1-1/n} \frac{x}{2}\mathrm{d}x = 1$$

$$\mathrm{E}T_n^2 = n^2 \cdot \frac{1}{n} + \int_{-1+1/n}^{1-1/n} \frac{x^2}{2}\mathrm{d}x$$

$$= n + (1-1/n)^3/3$$

亦即 $\mathrm{Var}(T_n) = n + (1-1/n)^3/3 - 1 \to \infty$，而 X 的方差 $\mathrm{Var}(X) = \dfrac{2^2}{12} = \dfrac{1}{3}$，可见 T_n 的方差没有收敛到其渐近分布的方差，而是趋于无穷.

在定理 2.14 的证明中，我们主要用到两点事实.其一是似然方程的渐近展开.其二是应用大数定律，由此可写出

$$\sqrt{n}(\hat{\theta} - \theta) - \sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln p(X_i;\theta)}{\partial\theta}I^{-1}(\theta) = o_p(1) \quad (2.35)$$

受此启发，只要(2.35)式成立，然后应用中心极限定理便可证得 $\hat{\theta}$ 的渐近正态性，因此，定理 2.14 可作多种形式的推广，比如推广到有限维向量参数.设 $\theta \in \Theta \subset \mathbf{R}^k$，如果可以由多变量 Taylor 展开式得到

$$\sqrt{n}(\hat{\theta} - \theta) - \frac{1}{\sqrt{n}}I^{-1}(\theta)\sum_{i=1}^{n}S_\theta(x_i) = o_p(1)$$

其中 S_θ 已在定义 2.8 中给出(它也称为记分向量)，则 $\hat{\theta}$ 是渐近正态的.

例 2.31 设 $X_1, \cdots, X_n$ 为独立同分布变量，

$$Pr(X_1=j)=\pi_j,\ j=1,\cdots,m.$$

因 $\sum_{j=1}^{m}\pi_j=1$，为去除约束条件，以 $\pi_m=1-\sum_{j=1}^{m-1}\pi_j$ 代入，记 $\boldsymbol{\theta}=(\pi_1,\cdots,\pi_{m-1})'$，则

$$\begin{aligned}\ln p(x_1;\boldsymbol{\theta})&=\sum_{j=1}^{m}I_{(x_1=j)}\ln\pi_j\\&=\sum_{j=1}^{m-1}I_{(x_1=j)}\ln\pi_j+I_{(x_1=m)}\ln\left(1-\sum_{j=1}^{m-1}\pi_j\right)\end{aligned}$$

$$\frac{\partial\ln p(x_1;\boldsymbol{\theta})}{\partial\pi_j}=\frac{I_{(x_1=j)}}{\pi_j}-\frac{I_{(x_1=m)}}{\pi_m},(j\neq m)$$

$$\frac{\partial^2\ln p(x_1;\boldsymbol{\theta})}{\partial\pi_i\partial\pi_j}=\begin{cases}-\left(\dfrac{I_{(x_1=j)}}{\pi_j^2}+\dfrac{I_{(x_1=m)}}{\pi_m^2}\right), & i=j(\neq m)\\ -\dfrac{I_{(x_1=m)}}{\pi_m^2}, & i\neq j(\neq m)\end{cases}$$

$$\begin{aligned}I_{ij}(\boldsymbol{\theta})&=-\mathrm{E}\left(\frac{\partial^2\ln p(X_1;\boldsymbol{\theta})}{\partial\pi_i\partial\pi_j}\right)\\&=\frac{1}{\pi_m}+\frac{\delta_{ij}}{\pi_j},\quad i,j=1,\cdots,m-1\end{aligned}$$

其中 $\delta_{ij}=1$，如果 $i=j$；$\delta_{ij}=0$，如果 $i\neq j$. 用矩阵形式表示有

$$\boldsymbol{I}(\boldsymbol{\theta})=\mathrm{diag}\left(\frac{1}{\pi_1},\cdots,\frac{1}{\pi_{m-1}}\right)+\frac{1}{\pi_m}\mathbf{1}_{m-1}\mathbf{1}'_{m-1}$$

记 $A=\mathrm{diag}\left(\frac{1}{\pi_1},\cdots,\frac{1}{\pi_{m-1}}\right)$，$\boldsymbol{U}=\mathbf{1}_{m-1}/\sqrt{\pi_m}$，则

$$\begin{aligned}\boldsymbol{I}^{-1}(\boldsymbol{\theta})&=\boldsymbol{A}^{-1}-\frac{\boldsymbol{A}^{-1}\boldsymbol{U}\boldsymbol{U}'\boldsymbol{A}^{-1}}{1+\boldsymbol{U}'\boldsymbol{A}^{-1}\boldsymbol{U}}\\&=\mathrm{diag}(\pi_1,\cdots,\pi_{m-1})-\frac{\boldsymbol{\theta}\boldsymbol{\theta}'/\pi_m}{1+\sum_{j=1}^{m-1}\pi_j/\pi_m}\\&=\mathrm{diag}(\pi_1,\cdots,\pi_{m-1})-\boldsymbol{\theta}\boldsymbol{\theta}'\end{aligned}$$

若记 $N_j=\#\{x_i:x_i=j\}$，$j=1,\cdots,m$，则 $\boldsymbol{\theta}$ 的 MLE 为 $\hat{\pi}_j=\dfrac{N_j}{n}$，$j=1,\cdots,m-1$，于是有

$$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\xrightarrow{L}N(0,\Sigma)$$

其中 $\boldsymbol{\Sigma}=(\sigma_{ij})$，$\sigma_{ij}=\pi_j\delta_{ij}-\pi_i\pi_j$. 该结果与上节中的(2.22)式是一致的，事实上，此处的渐近方差正是其精确方差.

定理 2.14 还可以推广至其它场合，比如独立但不同分布情形(如回归分析)以及不独立情形(如 Markov 链)等，有兴趣的读者可参阅有关文献[8].

§2.5.3 渐近有效性

定理 2.14 表明，在一定的正则条件下，MLE 的渐近方差为 $[nI(\theta)]^{-1}$，回顾 2.3 节，$[nI(\theta)]^{-1}$正是由容量为 n 的样本得到的 θ 的无偏估计的方差下界，达到此 C-R 下界的无偏估计称为有效无偏估计. 因此，很自然地可以将达到 C-R 下界的渐近无偏的估计序列称为渐近有效的.

定义 2.13 设 T_n 是 $g(\theta)$的相合估计，$g(\theta)$可微，Fisher 信息存在，且

$$\sqrt{n}(T_n-g(\theta))\xrightarrow{L}N(0,\sigma^2(\theta))$$

则称

$$e(\theta,T_n)=\frac{[g'(\theta)]^2/I(\theta)\cdot n}{\sigma^2(\theta)}$$

为 T_n 的渐近效. 如果 $e(\theta,T_n)=1$，则称 T_n 是 $g(\theta)$的渐近有效的估计.

Fisher 早在 1925 年就猜想渐近分布为正态的相合估计的渐近方差不小于其 C-R 下界，因此用定义 2.13 来定义估计的渐近效有其合理性，我们已经看到，在一定条件下，MLE 是渐近有效的. 此处渐近效的概念比 2.3 节中定义的效更有意义，因为它(在正则场合)是可以达到 1 的.

在实际中，有时考虑一些虽不渐近有效但简便易用且效较高的估计是很有用的，因为渐近有效的估计可能很难求.

例 2.32 设 $X_1,\cdots,X_n$ 是来自 Cauchy 分布的一个样本，$p(x_1;\theta)=[\pi(1+(x_1-\theta)^2)]^{-1}$，$\theta\in\mathbf{R}$，则

$$l(\theta;x)=\ln L(\theta;x)=-\sum_{i=1}^{n}\ln[1+(x_i-\theta)^2]-n\ln\pi$$

似然方程为

$$\sum_{i=1}^{n}\frac{x_i-\theta}{1+(x_i-\theta)^2}=0$$

通分可看出这是 θ 的 $2n-1$ 次多项式方程，解它很困难，要借助计算机，而且，它可能有许多个解，我们要对每个解计算其似然函数加以比较才能找到 MLE，这是一个相当繁琐的过程.

由于 θ 是 Cauchy 分布的中位数，我们可用样本中位数 $m_{0.5}$ 估计 θ，该估计很容易计算，其渐近方差为 $\pi^2/(4n)$，那末它的效如何呢？对此，要计算该分布族的 Fisher 信息，

$$\begin{aligned}I_1(\theta)&=\mathrm{E}\left\{\frac{2(X-\theta)}{1+(X-\theta)^2}\right\}^2\\&=\int_{-\infty}^{\infty}\left[\frac{2(x-\theta)}{1+(x-\theta)^2}\right]^2\frac{1}{\pi[1+(x-\theta)^2]}\mathrm{d}x\\&=\int_{-\infty}^{\infty}\frac{4x^2}{\pi(1+x^2)^3}\mathrm{d}x\end{aligned}$$

令 $x=\tan t$，则

$$\begin{aligned}I_1(\theta)&=\frac{1}{\pi}\int_{-\pi/2}^{\pi/2}4\tan^2 t\sec^4 t\,\mathrm{d}t\\&=\frac{1}{\pi}\int_{-\pi/2}^{\pi/2}\sin^2(2t)\mathrm{d}t\\&=1/2\end{aligned}$$

故 $m_{0.5}$ 的渐近效为 $\frac{2}{\pi^2/4}=8/\pi^2$（约为 0.81），很不错.

§2.5.4 局限性

极大似然方法虽然应用很广，以至于有人认为一切统计推断都应从似然函数出发，似然函数 $L(\theta;x)$ 包含了 θ 的全部信息，然而这样的观点有其局限性. 下述一些例子说明了极大似然方法的某些不足.

例 2.33 假设我们关心的是一个由 N 个个体组成的总体，其第 j 个个体有一个指标值 μ_j，$\theta=\sum_{j=1}^{N}\mu_j$ 是我们感兴趣的参数（这在实际中是经常出现的，譬如，我们要了解上海市工人的平均收入）. 现设不放

回随机抽取 n 个个体，则

$$Pr(X_i = \mu_{j_i}, i = 1,\cdots,n) = n!(N-n)!/N!$$

显然，上述概率函数未包含 θ 的任何信息，亦即似然函数对估计 θ 是没有用的. 然而这并不能说样本不包含 θ 的信息，事实上，此处样本确实包含了 θ 的信息，譬如，$\hat{\theta}=N\overline{X}$ 就是 θ 的一个不错的估计.

例 2.34 设 $X_1,\cdots,X_n$ 是来自 $U(0,\theta)$ 的一个样本，则

$$L(\theta;x) = \begin{cases} \theta^{-n}, & 0 < X_{(1)} < X_{(n)} < \theta \\ 0, & \text{其它} \end{cases}$$

显然，$\hat{\theta}=X_{(n)}$ 是 θ 唯一的 MLE，然而直观上看，$X_{(n)}$ 不是一个好的估计，因为 $P_\theta(X_{(n)}<\theta)=1$，$X_{(n)}$ 总是低估 θ，事实上，我们可以找到一个一致优于 $X_{(n)}$ 的估计，譬如，如果以均方误差作为衡量标准，则 θ 的 UMVUE $\frac{n+1}{n}X_{(n)}$ 一致优于 $X_{(n)}$（进一步，$\frac{n+2}{n+1}X_{(n)}$ 在形如 $\alpha X_{(n)}$ 的估计中均方误差达到最小）.

本例也说明，极大似然估计并不一定都通过求导数获得.

例 2.35 设 $X_1,\cdots,X_n$ 是来自 $U(\theta,\theta+1)$ 的一个样本，

$$L(\theta;x) = \begin{cases} 1, & X_{(n)} - 1 < \theta < X_{(1)} \\ 0, & \text{其它} \end{cases}$$

因此，任意一个介于 $X_{(n)}-1$ 与 $X_{(1)}$ 之间的数都是 θ 的一个 MLE，如 $X_{(n)}-1$，$X_{(1)}$ 等等，这也等于是说，极大似然方法未能唯一确定 θ 的点估计.

另外，从统计决策理论（见第五章）来看，任何统计推断都应依赖损失函数，而极大似然方法未曾考虑到损失函数，难免让人觉得不满意. 综合上述，极大似然方法虽有一些优良性，但也有其局限性.

§2.6 最小二乘估计

§2.6.1 最小二乘估计

最小二乘法是一种常用的估计方法，最常见于线性模型. 下面考虑

Gauss-Markov 模型：

$$\mathrm{E}\boldsymbol{Y}=\boldsymbol{X\beta};\mathrm{Var}(\boldsymbol{Y})=\sigma^2\boldsymbol{I}_n$$

其中 $\boldsymbol{Y}$ 为 $n\times1$ 维观测向量，$\boldsymbol{X}$ 为已知的 $n\times p(p\leqslant n)$ 阶设计矩阵，$\boldsymbol{\beta}$ 为 $p\times1$ 维未知参数，σ^2 未知，$\boldsymbol{I}_n$ 为 n 阶单位阵. 此模型也称为独立观测线性模型，通常记为 $(\boldsymbol{Y},\boldsymbol{X\beta},\sigma^2\boldsymbol{I}_n)$.

为估计 β，一般采用最小二乘法.

定义 2.14 在 $(\boldsymbol{Y},\boldsymbol{X\beta},\sigma^2\boldsymbol{I}_n)$ 中，如果

$$(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})=\min_{\beta}(\boldsymbol{Y}-\boldsymbol{X\beta})'(\boldsymbol{Y}-\boldsymbol{X\beta}) \quad (2.36)$$

则称 $\hat{\boldsymbol{\beta}}$ 为 $\boldsymbol{\beta}$ 的最小二乘估计(Least Squares Estimate)，简称 LSE.

记 $Q(\boldsymbol{\beta})=(\boldsymbol{Y}-\boldsymbol{X\beta})'(\boldsymbol{Y}-\boldsymbol{X\beta})$，求 LSE 等价于求 $Q(\boldsymbol{\beta})$ 的最小值. 利用矩阵求导，有

$$\frac{\partial \boldsymbol{Y}'\boldsymbol{X\beta}}{\partial\boldsymbol{\beta}}=\boldsymbol{X}'\boldsymbol{Y},\quad \frac{\partial\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta}}{\partial\boldsymbol{\beta}}=2\boldsymbol{X}'\boldsymbol{X\beta}$$

从而

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=-2\boldsymbol{X}'\boldsymbol{Y}+2\boldsymbol{X}'\boldsymbol{X\beta}$$

令此导数等于 0，给出

$$\boldsymbol{X}'\boldsymbol{X\beta}=\boldsymbol{X}'\boldsymbol{Y} \quad (2.37)$$

这就是所谓的正规方程，它在最小二乘估计理论中起着重要作用.

简单起见，设 X 是列满秩阵，则正规方程(2.37)的解唯一，它就是 β 的 LSE

$$\hat{\boldsymbol{\beta}}_{LS}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \quad (2.38)$$

显然，$\hat{\boldsymbol{\beta}}_{LS}$ 是 $\boldsymbol{Y}$ 的线性函数.

定理 2.15 (1)若 $\mathrm{E}(\boldsymbol{Y})=\boldsymbol{X\beta}$，则(2.38)式给出的 $\hat{\boldsymbol{\beta}}_{LS}$ 是 $\boldsymbol{\beta}$ 的无偏估计.

(2)若 $\mathrm{E}(\boldsymbol{Y})=\boldsymbol{X\beta}$，$\mathrm{Var}(\boldsymbol{Y})=\sigma^2\boldsymbol{I}_n$，$\boldsymbol{X}$ 为列满秩阵，则(2.38)式给出的 $\hat{\boldsymbol{\beta}}_{LS}$ 的方差为 $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

证明：(1) $\mathrm{E}(\hat{\boldsymbol{\beta}}_{LS})=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{E}(\boldsymbol{Y})=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X\beta}=\boldsymbol{\beta}$.

(2)由(1)，$\hat{\boldsymbol{\beta}}_{LS}$是$\boldsymbol{\beta}$的无偏估计，故

$$
\begin{aligned}
&\mathrm{Var}(\hat{\boldsymbol{\beta}}_{LS}) = \mathrm{E}(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})' \\
&= \mathrm{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{E}(\boldsymbol{Y})] \cdot \\
&\quad [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{E}(\boldsymbol{Y})]' \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{E}[\boldsymbol{Y} - \mathrm{E}(\boldsymbol{Y})][\boldsymbol{Y} - \mathrm{E}(\boldsymbol{Y})]'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}
\end{aligned}
$$

证毕.

例 2.36　设某因子 x 对指标 y 有线性影响关系，即

$$
y = \beta_0 + \beta_1 x + \varepsilon,
$$

$$
\mathrm{E}(\varepsilon) = 0, \quad \mathrm{Var}(\varepsilon) = \sigma^2
$$

为估计 β_0, β_1，取因子 x 的 n 个水平 $X_1, \cdots, X_n$ 进行观测，得到 y 的 n 个独立观测值，此便是 Gauss-Markov 模型 $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$，其中

$$
\boldsymbol{Y}' = (Y_1, \cdots, Y_n), \boldsymbol{X}' = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{pmatrix}
$$

$$
\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}
$$

易知，只要 x 的 n 个水平不全相同，$\boldsymbol{X}$ 为列满秩阵. 不难算出

$$
\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix}
$$

$$
(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{1}{n\sum_{i=1}^{n}(X_i - \bar{X})^2}\begin{pmatrix} \sum X_i^2 & \sum X_i \\ -\sum X_i & n \end{pmatrix}
$$

由(2.38)式，$\boldsymbol{\beta} = (\beta_0, \beta_1)'$的 LSE 为

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{LS} &= \frac{1}{n\sum_{i=1}^{n}(X_i - \bar{X})^2}\begin{pmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{pmatrix}\begin{pmatrix} \sum Y_i \\ \sum (X_i Y_i) \end{pmatrix} \\
&= \frac{1}{n\sum_{i=1}^{n}(X_i - \bar{X})^2}\begin{pmatrix} [\sum X_i^2 \sum Y_i - (\sum X_i)^2 \sum Y_i] + [(\sum X_i)^2 \sum Y_i - \sum X_i \sum (X_i Y_i)] \\ n\sum (X_i Y_i) - \sum X_i \sum Y_i \end{pmatrix} \\
&= \begin{pmatrix} \bar{Y} - \bar{X}\sum[(X_i - \bar{X})(Y_i - \bar{Y})]/\sum (X_i - \bar{X})^2 \\ \sum[(X_i - \bar{X})(Y_i - \bar{Y})]/\sum (X_i - \bar{X})^2 \end{pmatrix}
\end{aligned}
$$

由定理 2.15 知，$\hat{\boldsymbol{\beta}}_{LS}$是$\boldsymbol{\beta}$的无偏估计，且

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{LS}) = \frac{\sigma^2}{n\sum_{i=1}^{n}(X_i - \overline{X})^2}\begin{pmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{pmatrix}$$

特别，$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\dfrac{\sum X_i}{n\sum(X_i - \overline{X})^2}$，如果在试验条件允许下，或经过某种数据变换，可以使得 $\sum X_i = 0$，则 LSE $\hat{\beta}_0$ 与$\hat{\beta}_1$ 不相关. 这不仅使计算简化，而且使 LSE $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 间的干扰减少了. 这是试验设计专门研究的内容，此处从略.

§2.6.2 最好线性无偏估计

在$(\boldsymbol{Y}, \boldsymbol{X\beta}, \sigma^2\boldsymbol{I}_n)$中，若 $\boldsymbol{X}$ 列满秩，则 LSE 还在某种意义下具有最优性.

定义 2.15 若估计量是观测值的线性函数，则称它为线性估计. 设 $\hat{\theta}$ 是参数 θ 的线性无偏估计，如果对 θ 的任意一个线性无偏估计 θ^*，有

$$\mathrm{Var}(\theta^*) \geqslant \mathrm{Var}(\hat{\theta}), \quad \forall\, \theta \in \Theta$$

则 $\hat{\theta}$ 称为 θ 的最好线性无偏估计(Best Linear Unbiased Estimate)，简称 BLUE.

如果 θ 是 k 维向量参数，BLUE 的定义是类似的，其中 $\mathrm{Var}(\theta^*) \geqslant \mathrm{Var}(\hat{\theta})$表示 $\mathrm{Var}(\theta^*) - \mathrm{Var}(\hat{\theta})$是非负定阵.

在定义 2.15 的标准下，LSE 是 BLUE.

定理 2.16 设 $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{X\beta}$, $\mathrm{Var}(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}_n$, $\boldsymbol{X}$ 为列满秩阵，则 LSE $\hat{\boldsymbol{\beta}}_{LS}$是 $\boldsymbol{\beta}$ 唯一的 BLUE.

证明：前面已证$\hat{\boldsymbol{\beta}}_{LS}$是 $\boldsymbol{\beta}$ 的无偏估计，而线性是显然的，下证$\hat{\boldsymbol{\beta}}_{LS}$的方差在 $\boldsymbol{\beta}$ 的线性无偏估计类中是最小的. 设 $\varphi(\boldsymbol{Y}) = \boldsymbol{AY}$ 是 $\boldsymbol{\beta}$ 的任一个线性无偏估计，则应有

$$\mathrm{E}(\boldsymbol{AY}) = \boldsymbol{A}\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{AX\beta} = \boldsymbol{\beta}, \quad \forall\, \boldsymbol{\beta} \in \boldsymbol{R}^p$$

于是有 $\boldsymbol{AX} = \boldsymbol{I}_p$，其方差为

$$\mathrm{Var}(\boldsymbol{AY}) = \mathrm{E}[(\boldsymbol{AY} - \hat{\boldsymbol{\beta}}_{LS} + \hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})(\boldsymbol{AY} - \hat{\boldsymbol{\beta}}_{LS} + \hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})']$$

$$= \mathrm{E}[(AY - \hat{\beta}_{LS})(AY - \hat{\beta}_{LS})'] + \mathrm{Var}(\hat{\beta}_{LS}) -$$
$$\mathrm{E}[(AY - \hat{\beta}_{LS})(\hat{\beta}_{LS} - \beta)'] - \mathrm{E}[(\hat{\beta}_{LS} - \beta)(AY - \hat{\beta}_{LS})']$$

又因为 $\mathrm{E}(AY-\hat{\beta}_{LS})=0$,所以

$$\begin{aligned}
&\mathrm{E}[(AY - \hat{\beta}_{LS})(\hat{\beta}_{LS} - \beta)'] \\
&= \mathrm{Cov}(AY - \hat{\beta}_{LS}, \hat{\beta}_{LS}) \\
&= \mathrm{Cov}([A - (X'X)^{-1}X']Y, (X'X)^{-1}X'Y) \\
&= (A - (X'X)^{-1}X')\mathrm{Cov}(Y,Y)((X'X)^{-1}X')' \\
&= \sigma^2(AX - (X'X)^{-1}X'X)(X'X)^{-1} = 0
\end{aligned}$$

由此即有 $\mathrm{Var}(AY)\geqslant\mathrm{Var}(\hat{\beta}_{LS})$,且等号成立的充要条件是 $\mathrm{E}[(AY-\hat{\beta}_{LS})(Y-\hat{\beta}_{LS})']=0$,即 $AY=\hat{\beta}_{LS}$,a. s. 证明完成.

定理 2.17 在 $(Y, X\beta, \sigma^2 I_n)$ 中,若 X 列满秩,则 $\forall\ \alpha\in\mathbf{R}^p$,$\alpha'\hat{\beta}_{LS}$ 是 $\alpha'\beta$ 唯一的 BLUE.

证明:$\alpha'\hat{\beta}_{LS}$ 是 $\alpha'\beta$ 的线性无偏估计是显然的,其最优性可以仿照定理 2.16 的证明进行,此处从略.

在例 2.36 中,由定理 2.16,LSE $\hat{\beta}_{LS}$ 是 β 的 BLUE. 据定理 2.17,若令 $\alpha=(1,0)$ 或 $\alpha=(0,1)$,则 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 分别是 β_0 和 β_1 的 BLUE. 若令 $\alpha=(1,x_0)$,则 $\hat{\beta}_0+\hat{\beta}_1x_0$ 还是 $\beta_0+\beta_1x_0$ 的 BLUE. 这些性质都表明 LSE 是一个很好的估计.

在线性模型中,σ^2 的估计也是一个重要内容. 由于量纲的关系,σ^2 的估计不可能用样本的线性函数给出,但可以用样本的二次函数给出. 为此,称

$$R_0^2 = (Y - X\hat{\beta}_{LS})'(Y - X\hat{\beta}_{LS})$$

为偏差平方和,它就是(2.36)式中目标函数 $Q(\beta)$ 的最小值,即 $R_0^2=Q(\hat{\beta}_{LS})$. 若 X 为 p 维列满秩矩阵,则可以证明

$$\hat{\sigma}^2 = R_0^2/(n-p) \tag{2.39}$$

是 σ^2 的无偏估计，其证明留给读者自己进行. 由于 σ^2 的这个估计是基于 $\boldsymbol{\beta}$ 的 LSE 导出的，因而也常将之称为 σ^2 的 LSE，关于它的优良性可查阅有关文献[9].

当 $\boldsymbol{Y}$ 服从正态分布时，上述 LSE,BLUE 与 UMVUE 重合.

定理 2.18 在 $(\boldsymbol{Y},\boldsymbol{X\beta},\sigma^2\boldsymbol{I}_n)$ 中，若假定 $\boldsymbol{Y}\sim N_n(\boldsymbol{X\beta},\sigma^2\boldsymbol{I}_n)$，$\boldsymbol{X}$ 为列满秩阵，则 (2.38) 式中的 $\hat{\boldsymbol{\beta}}_{LS}$ 和 (2.39) 式中的 $\hat{\sigma}^2$ 分别是 $\boldsymbol{\beta}$ 和 σ^2 的 UMVUE. 另外，$\forall\ \boldsymbol{\alpha}\in\mathbf{R}^p$，$\boldsymbol{\alpha}'\hat{\boldsymbol{\beta}}_{LS}$ 也是 $\boldsymbol{\alpha}'\hat{\boldsymbol{\beta}}$ 的 UMVUE.

证明：$Y_1,\cdots,Y_n$ 的联合密度函数为

$$p(y_1,\cdots,y_n;\boldsymbol{\beta},\sigma^2)=\left(\sqrt{2\pi}\sigma\right)^{-n}\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{Y}-\boldsymbol{X\beta})'(\boldsymbol{Y}-\boldsymbol{X\beta})\right\}$$

$$=Q(\theta)\exp\{\theta_1T_1(\boldsymbol{Y})+\theta_2T_2(\boldsymbol{Y})\}$$

其中 $\theta_1=-\frac{1}{2\sigma^2}$，$\theta_2=\frac{\boldsymbol{\beta}}{\sigma^2}$，$T_1(\boldsymbol{Y})=\boldsymbol{Y}'\boldsymbol{Y}$，$T_2(\boldsymbol{Y})=\boldsymbol{X}'\boldsymbol{Y}$，$Q(\theta)=\left(\sqrt{2\pi}\sigma\right)^{-n}\cdot\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{X\beta})'\boldsymbol{X\beta}\right\}$. 由指数族性质知 $T_1(Y)$ 和 $T_2(Y)$ 为完备充分统计量，而

$$\hat{\boldsymbol{\beta}}_{LS}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}=(\boldsymbol{X}'\boldsymbol{X})^{-1}T_2(\boldsymbol{Y})$$

$$\hat{\sigma}^2=\frac{1}{n-p}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{LS})'(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{LS})$$

$$=\frac{1}{n-p}(\boldsymbol{Y}'\boldsymbol{Y}-2\boldsymbol{Y}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_{LS}+\hat{\boldsymbol{\beta}}_{LS}\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_{LS})$$

$$=\frac{1}{n-p}(T_1(\boldsymbol{Y})-T'_2(\boldsymbol{Y})(\boldsymbol{X}'\boldsymbol{X})^{-1}T_2(\boldsymbol{Y}))$$

这表明 $\hat{\boldsymbol{\beta}}_{LS}$ 和 $\hat{\sigma}^2$ 都是完备充分统计量的函数，又因它们分别是 $\boldsymbol{\beta}$ 和 σ^2 的无偏估计，由定理 2.3 知，$\hat{\boldsymbol{\beta}}_{LS}$ 和 $\hat{\sigma}^2$ 分别是 $\boldsymbol{\beta}$ 和 σ^2 唯一的 UMVUE. 类似可证 $\boldsymbol{\alpha}'\hat{\boldsymbol{\beta}}_{LS}$ 是 $\boldsymbol{\alpha}'\boldsymbol{\beta}$ 的 UMVUE. 证毕.

以上讨论均假定 $\boldsymbol{X}$ 为列满秩阵，倘 $\boldsymbol{X}$ 不是列满秩阵，也有相当令人满意的结论[9].

§2.6.3 加权最小二乘估计

在 $(\boldsymbol{Y},\boldsymbol{X\beta},\sigma^2\boldsymbol{I}_n)$ 中，如果条件 $\mathrm{Var}(\boldsymbol{Y})=\sigma^2\boldsymbol{I}_n$ 不满足，那末最小二乘

估计是否还具有上述 BLUE 等性质呢？假如不具有，那又如何改进呢？下面我们就讨论这个问题.

首先指出，若 $\mathrm{Var}(\boldsymbol{Y})=\sigma^2\boldsymbol{I}_n$ 不满足但 $\mathrm{E}(\boldsymbol{Y})=\boldsymbol{X}\boldsymbol{\beta}$ 成立时，最小二乘法仍然可以使用，算得的 LSE(2.38)也还是 $\boldsymbol{\beta}$ 的线性无偏估计，这是因为使用最小二乘法和无偏性只涉及一阶矩，并不需要二阶矩具有什么性质. 然而，BLUE 就涉及到二阶矩性质，此时，(2.38)一般不再是 $\boldsymbol{\beta}$ 的 BLUE.

在$(\boldsymbol{Y},\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n)$中，如果将条件 $\mathrm{Var}(\boldsymbol{Y})=\sigma^2\boldsymbol{I}_n$ 改为 $\mathrm{Var}(\boldsymbol{Y})=\sigma^2\boldsymbol{G}$，$\boldsymbol{G}$ 为已知正定阵，则形成所谓的广义 Gauss-Markov 模型. 对此模型，因 $\boldsymbol{G}>0$，存在 n 阶非奇异对称阵 $\boldsymbol{B}$，使 $\boldsymbol{G}=\boldsymbol{B}^2$. 令 $\widetilde{\boldsymbol{Y}}=\boldsymbol{B}^{-1}\boldsymbol{Y}$，$\widetilde{\boldsymbol{X}}=\boldsymbol{B}^{-1}\boldsymbol{X}$，则

$$\mathrm{E}\widetilde{\boldsymbol{Y}}=\boldsymbol{B}^{-1}\mathrm{E}\boldsymbol{Y}=\boldsymbol{B}^{-1}\boldsymbol{X}\boldsymbol{\beta}=\widetilde{\boldsymbol{X}}\boldsymbol{\beta}$$

$$\mathrm{Var}(\widetilde{\boldsymbol{Y}})=\boldsymbol{B}^{-1}\mathrm{Var}(\boldsymbol{Y})\boldsymbol{B}^{-1}=\sigma^2\boldsymbol{I}_n$$

由此，$(\widetilde{\boldsymbol{Y}},\widetilde{\boldsymbol{X}}\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n)$是一个 Gauss-Markov 模型，由该模型得到的 LSE

$$\widetilde{\boldsymbol{\beta}}=(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}=(\boldsymbol{X}'\boldsymbol{G}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{G}^{-1}\boldsymbol{Y} \tag{2.40}$$

称为 $\boldsymbol{\beta}$ 的加权最小二乘估计，由定理 2.16，它是 $\boldsymbol{\beta}$ 的 BLUE.

加权最小二乘估计应用很广，尤其是对截尾样本.

考虑一类常用的位置-尺度分布族，其分布函数有形式 $F\left(\dfrac{y-\mu}{\sigma}\right)$，密度函数有形式$\dfrac{1}{\sigma}f\left(\dfrac{y-\mu}{\sigma}\right)$，其中 $\mu(-\infty<\mu<\infty)$称为位置参数，$\sigma>0$ 称为尺度参数. 设 $Y_1,\cdots,Y_n$ 是来自 $F\left(\dfrac{y-\mu}{\sigma}\right)$的一个样本，$F(\cdot)$为已知的分布函数且二阶矩存在，我们要估计 μ 和 $\sigma(\sigma>0)$. 对某些 F（譬如正态分布），有众知的一些方法和结果. 此处我们考虑这样一类估计，它们是次序统计量的线性函数.

设 $Y_{(1)}\leqslant\cdots\leqslant Y_{(r)}$为观测到的前 r 个次序统计量. 令 $X_{(i)}=(Y_{(i)}-\mu)/\sigma$，$i=1,\cdots,r$，则 $X_{(1)}\leqslant\cdots\leqslant X_{(r)}$相当于抽自 $F(x)$的容量为 n 的前 r 个次序样本. 记

$$\mathrm{E}X_{(i)}=\alpha_i,\quad i=1,\cdots,r$$

$$\mathrm{Cov}(X_{(i)},X_{(j)})=v_{ij},\quad 1\leqslant i,j\leqslant r$$

其中 α_i, v_{ij} 只依赖于 n, r 和 F 而与 μ, σ 无关. 由于

$$Y_{(i)} = \mu + \sigma X_{(i)} = \mu + \sigma\alpha_i + \varepsilon_i$$

其中 $\varepsilon_i = \sigma(X_{(i)} - \alpha_i)$. 记 $\boldsymbol{Y}' = (Y_{(1)}, \cdots, Y_{(r)})$, $\boldsymbol{\alpha}' = (\alpha_1, \cdots, \alpha_r)$, 用矩阵表示, 有

$$\mathrm{E}\boldsymbol{Y} = (\mathbf{1}_r, \boldsymbol{\alpha})\begin{pmatrix}\mu \\ \sigma\end{pmatrix}$$

$$\mathrm{Var}(\boldsymbol{Y}) = \sigma^2 \boldsymbol{V} = \sigma^2 (v_{ij})_{r\times r}$$

其中 $\mathbf{1}_r$ 表示全部由元素 1 组成的 r 维列向量. 这是广义 Gauss-Markov 模型, 利用(2.40)式可求出 μ 和 σ 的 BLUE 为

$$\begin{pmatrix}\hat{\mu} \\ \hat{\sigma}\end{pmatrix} = \begin{pmatrix}\mathbf{1}'_r \boldsymbol{V}^{-1}\mathbf{1}_r & \boldsymbol{\alpha}'\boldsymbol{V}^{-1}\mathbf{1}_r \\ \boldsymbol{\alpha}'\boldsymbol{V}^{-1}\mathbf{1}_r & \boldsymbol{\alpha}'\boldsymbol{V}^{-1}\boldsymbol{\alpha}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{1}'_r \\ \boldsymbol{\alpha}'\end{pmatrix}\boldsymbol{V}^{-1}\boldsymbol{Y}$$

$$\triangleq \begin{pmatrix}\boldsymbol{l}'_1\boldsymbol{Y} \\ \boldsymbol{l}'_2\boldsymbol{Y}\end{pmatrix} \tag{2.41}$$

其协差阵为

$$\mathrm{Var}\begin{pmatrix}\hat{\mu} \\ \hat{\sigma}\end{pmatrix} = \sigma^2\begin{pmatrix}\boldsymbol{l}'_1\boldsymbol{V}\boldsymbol{l}_1 & \boldsymbol{l}'_1\boldsymbol{V}\boldsymbol{l}_2 \\ \boldsymbol{l}'_1\boldsymbol{V}\boldsymbol{l}_2 & \boldsymbol{l}'_2\boldsymbol{V}\boldsymbol{l}_2\end{pmatrix} \tag{2.42}$$

该估计方法有一个明显的优点: 不论 n 个样品中被观测到的样品个数是多少(当然不能少于 2), 上述过程均可类似进行. 该性质使得它在可靠性中有着广泛的应用, 譬如, 当 $F(y)$ 是 I 型极值分布 $1-\exp\{-\exp(y)\}$ 时, 对大量的 n 和 r, (2.41)式中的系数 $\boldsymbol{l}_1, \boldsymbol{l}_2$ 已编制成表, 可直接查用, 这使得该估计方法的运用更显方便[6].

例 2.37 Weibull 分布是一类常见的产品寿命分布. 设某电子产品的寿命 T 服从 Weibull 分布, 其分布函数为

$$Pr(T < t) = 1 - \exp\left\{-\left(\frac{t}{\eta}\right)^m\right\}, \quad t > 0$$

其中 $m > 0$ 称为形状参数, $\eta > 0$ 称为特征寿命.

令 $Y = \ln T$, 则 Y 服从 I 型极值分布, 其分布函数为

$$F(y) = Pr(Y < y) = 1 - \exp\{-\exp[(y - \mu)/\sigma]\}$$

其中 $\mu = \ln\eta$, $\sigma = 1/m$, μ 和 σ 分别是极值分布的位置参数和尺度参数.

为了对产品寿命特征进行分析, 投入 n 个产品样品进行寿命试验,

当有 r 个样品失效时试验停止，得到 r 个次序观测值 $t_{(1)}<t_{(2)}<\cdots<t_{(r)}$，相应地有 $y_{(1)}<y_{(2)}<\cdots<y_{(r)}$. 作变换 $x_{(i)}=(y_{(i)}-\mu)/\sigma$，则 $x_{(1)}<\cdots<x_{(r)}$ 可看作为来自标准极值分布 $1-\exp\{-\exp(x)\}$ 的容量为 n 的样本的前 r 个次序观测值. 显然，诸 $\mathrm{E}(X_{(i)})$ 和 $\mathrm{Cov}(X_{(i)},X_{(j)})$，$i,j=1,\cdots,r$ 只与 n 和 r 有关，而不依赖于其它参数. 由(2.41)式，可写出 σ 和 μ 的 BLUE

$$\hat{\sigma}=\sum_{j=1}^{r}C(n,r,j)y(j) \tag{2.43}$$

$$\hat{\mu}=\sum_{j=1}^{r}D(n,r,j)y(j) \tag{2.44}$$

其中 $C(n,r,j)$ 称为 σ 的最好线性无偏估计系数，$D(n,r,j)$ 称为 μ 的最好线性无偏估计系数. 对 $n=2,3,\cdots,25$，$r=2,3,\cdots,n$ 和 $j=1,\cdots,r$，它们的值可在文献[6]中的表一中查到. 譬如，若投人 8 个样品试验到 5 个样品失效时试验停止，记录下 5 个失效时间 $t_{(1)}<\cdots<t_{(5)}$(见表 2.2)，由文献[6]中表一查到诸 $C(n,r,j)$ 和 $D(n,r,j)$，由(2.43)，(2.44)式可计算 σ 和 μ 的 BLUE(见表 2.2)为 $\hat{\sigma}=1.4611$，$\hat{\mu}=4.5153$.

表 2.2 极值分布中参数的 BLUE 计算表($n=8,r=5$)

j	$t_{(j)}$	$y_{(j)}=\ln t_{(j)}$	$C(n,r,j)$	$C(n,r,j)y_{(j)}$	$D(n,r,j)$	$D(n,r,j)y_{(j)}$
1	2.5	0.9163	−0.2172	−0.1990	−0.0781	−0.0716
2	7.5	2.0149	−0.2128	−0.4288	−0.0474	−0.0955
3	17.0	2.8332	−0.1803	−0.5108	−0.0001	−0.0003
4	44.5	3.7955	−0.1225	−0.4649	0.0637	0.2418
5	65.5	4.1821	0.7328	3.0646	1.0619	4.4409
求和				1.4611		4.5153

§2.7 同变估计

§2.7.1 有偏估计

在2.1节中我们已指出，均方误差是评价一个点估计好坏的常用标准，而使均方误差一致达到最小的点估计是不存在的，于是，人们考虑在某些估计类中寻找使均方误差达到最小的点估计，无偏估计是一类最常用的点估计. 然而，必须指出，从均方误差角度来看，无偏估计并不一定就好. 假设 $\hat{\theta}(X)$ 是 θ 的无偏估计，现考虑形如 $a\hat{\theta}(X)$ 的估计，$a>0$ 是常数. 除非 $a=1$，$a\hat{\theta}(X)$ 是有偏估计.

$$\begin{aligned}\mathrm{MSE}(a\hat{\theta}(X)) &= \mathrm{E}(a\hat{\theta}(X)-\theta)^2\\ &= a^2\mathrm{E}(\hat{\theta}(X)-\theta)^2+(1-a)^2\theta^2\end{aligned}$$

由简单的求导知，当

$$a=\theta^2/(\theta^2+\mathrm{Var}(\hat{\theta})) \tag{2.45}$$

时，均方误差达最小. 当然，(2.45)式中的 a 可能依赖于 θ，$a\hat{\theta}(X)$ 不一定能作为估计，但由此我们也可看出，可能存在一个估计，它的均方误差比无偏估计的均方误差一致地小.

例 2.38 设 $X_1,\cdots,X_n$ 是来自 $N(\mu,\sigma^2)$ 的一个样本，现要估计 σ^2，我们知道其 UMVUE 是(2.6)式的 s^2，s^2 是 σ^2 的无偏估计类中使均方误差达到最小的. 现考虑形如 as^2 的估计，由于

$$\mathrm{Var}(s^2)=2\sigma^4/(n-1)$$

由(2.45)式知，使均方误差达最小的 a 是

$$a=(\sigma^2)^2/[(\sigma^2)^2+2\sigma^4/(n-1)]=(n-1)/(n+1)$$

此时，$as^2=\frac{1}{n+1}\sum(X_i-\overline{X})^2\triangleq\widetilde{\sigma}^2$. 可验算

$$\mathrm{MSE}(\widetilde{\sigma}^2)=\frac{2\sigma^2}{n+1}<\frac{2\sigma^2}{n-1}=\mathrm{MSE}(s^2)$$

由此，在均方误差意义下，有一个 σ^2 的有偏估计，它比所有 σ^2 的无偏估

计都来得好.

位置-尺度分布族 $F\left(\frac{x-\mu}{\sigma}\right)$ 是一类常见的分布族，μ 和 σ 的一个常用估计是 BLUE，我们在本节最后将指出，在均方误差的意义下，最好线性同变估计将优于 BLUE. 下面首先介绍同变估计.

§2.7.2　同变估计

在一些参数估计问题中，要求寻找的估计量在样本作某种特定变换下保持某种统计性质是很自然的. 譬如，若对随机变量 X 作一个位移变换：$X\to X+c$，因

$$E(X+c)=E(X)+c$$
$$\mathrm{Var}(X+c)=\mathrm{Var}(X)$$

现设有 X 的 n 个独立观测值 $X_1,\cdots,X_n$，如果用 $\hat{\mu}(X_1,\cdots,X_n)$ 和 $\hat{\sigma}^2(X_1,\cdots,X_n)$ 分别估计 X 的均值和方差，自然地可要求

$$\hat{\mu}(X_1+c,\cdots,X_n+c)=\hat{\mu}(X_1,\cdots,X_n)+c$$
$$\hat{\sigma}^2(X_1+c,\cdots,X_n+c)=\hat{\sigma}^2(X_1,\cdots,X_n)$$

这种要求是合理的，也是符合人们的实践经验的. 对样本作一个位移变换相当于改变计算原点，改变前后的数学期望的估计值应仅相差一个位移量，而方差估计值则不应有任何变化. 假如我们在选择估计量时不考虑这些要求，有时会觉得不自然，不合理. 满足这些要求的估计量就称为位移变换下的同变估计. 比如，样本均值 $\overline{X}$ 和样本方差 s_n^2 在位移变换下分别是总体均值和总体方差的同变估计.

显然，一个估计量是否是同变估计取决于采用什么变换. 除了上述位移变换以外，常考虑的变换还有尺度变换和线性变换. 若对随机变量 X 作一个尺度变换：$X\to\alpha X$，$\alpha>0$ 为常数，因

$$E(\alpha X)=\alpha E(X),\quad \mathrm{Var}(\alpha X)=\alpha^2\mathrm{Var}(X)$$

于是，在尺度变换下，总体均值 μ 和总体标准差 σ 的同变估计应满足

$$\hat{\mu}(\alpha X_1,\cdots,\alpha X_n)=\alpha\hat{\mu}(X_1,\cdots,X_n)$$
$$\hat{\sigma}(\alpha X_1,\cdots,\alpha X_n)=\alpha\hat{\sigma}(X_1,\cdots,X_n)$$

又如，考虑线性变换：$X\to\alpha X+c$，其中 $\alpha>0$ 和 c 是常数，由数学期望和

方差的性质可知,线性变换下总体均值 μ 和标准差 σ 的同变估计应满足

$$\hat{\mu}(aX_1 + c, \cdots, aX_n + c) = a\hat{\mu}(X_1, \cdots, X_n) + c$$

$$\hat{\sigma}(aX_1 + c, \cdots, aX_n + c) = \check{a}\hat{\sigma}(X_1, \cdots, X_n)$$

容易看出,在线性变换下,样本均值和样本方差分别是总体均值和方差的同变估计.

综合上述,我们可以给出同变估计的描述性定义,其严格定义需由变换群和统计决策问题进行,可查阅有关文献[3].

定义 2.16 设 $\hat{\theta}(X_1, \cdots, X_n)$是 θ 的一个估计量,如果在样本作某种特定变换下,估计量 $\hat{\theta}$ 具有某种相应的性质,则称 $\hat{\theta}$ 是在该变换下 θ 的同变估计. 在变换明确无误时,简称同变估计.

同变性与无偏性是对估计提出的两个不同的要求. 无偏性只有在多次重复使用下才有意义,而同变性是估计在每次使用时都应具备的性质. 有些同变估计是无偏的,但也有不少同变估计不具备无偏性. 事实上,有些有偏的同变估计在均方误差的意义下优于无偏估计.

在给定一类变换后,同变估计往往也不止一个. 譬如,在位移变换下,样本均值 $\overline{X}$ 和样本中位数 $m_{0.5}$都是总体均值的同变估计. 因此,在同变估计中也有一个如何选择的问题. 常用的选择标准是均方误差.

定义 2.17 设 $\hat{\theta}$ 是某特定变换下 θ 的一个同变估计,如果对 θ 的任一个此种变换下的同变估计 $\tilde{\theta}$,有

$$\mathrm{MSE}_\theta(\hat{\theta}) \leqslant \mathrm{MSE}_\theta(\tilde{\theta}), \quad \forall\, \theta \in \Theta$$

则称 $\hat{\theta}$ 是此种变换下 θ 的最优同变估计. 在变换明确无误时,简称 $\hat{\theta}$ 是 θ 的最优同变估计.

下面我们结合前面介绍的三类变换讨论最优同变估计.

§2.7.3 位置参数的同变估计

设 $X_1, \cdots, X_n$ 是来自分布密度函数为 $f(x-\theta)$的一个样本,其中 $f(\cdot)$已知,$\theta \in \mathbf{R}$ 是一个未知的位置参数. 假如 $\hat{\theta} = \hat{\theta}(X_1, \cdots, X_n)$是 θ 的位移变换下的同变估计,则对任意实数 c,应有

$$\hat{\theta}(X_1 + c, \cdots, X_n + c) = \hat{\theta}(X_1, \cdots, X_n) + c \tag{2.46}$$

令 $c=-X_1$，即有

$$\hat{\theta}(X_1,\cdots,X_n)=\hat{\theta}(0,X_2-X_1,\cdots,X_n-X_1)+X_1$$

记 $Y=(X_2-X_1,\cdots,X_n-X_1)$，上式可以写成

$$\hat{\theta}(X_1,\cdots,X_n)=X_1+h(Y) \tag{2.47}$$

这便是 θ 的同变估计的一般形式，而寻找最优同变估计就是要寻找 $h(\cdot)$ 使得(2.47)式的 $\hat{\theta}$ 的均方误差达到最小.

易见，$X_1-\theta$ 和 $h(Y)$ 的分布均与 θ 无关，故有

$$\begin{aligned}
&\text{MSE}_\theta(X_1+h(Y))=\text{E}_\theta(X_1+h(Y)-\theta)^2\\
&=\text{E}_0(X_1+h(Y))^2\\
&=\text{E}_0[X_1-\text{E}_0(X_1|Y)+\text{E}_0(X_1|Y)+h(Y)]^2\\
&=\text{E}_0(X_1-\text{E}_0(X_1|Y))^2+\text{E}(\text{E}_0(X_1|Y)+h(Y))^2
\end{aligned} \tag{2.48}$$

其中交叉乘积项

$$\begin{aligned}
&\text{E}_0[(X_1-\text{E}_0(X_1|Y))(\text{E}_0(X_1|Y)+h(Y))]\\
&=\text{E}_0\{\text{E}_0[(X_1-\text{E}_0(X_1|Y))(\text{E}_0(X_1|Y)+h(Y))|Y]\}\\
&=\text{E}_0\{(\text{E}_0(X_1|Y)+h(Y))\text{E}_0(X_1-\text{E}_0(X_1|Y)|Y)\}\\
&=0
\end{aligned}$$

由(2.48)知，当 $h(Y)=-\text{E}_0(X_1|Y)$ 时均方误差达到最小. 由此给出 θ 的最优同变估计为

$$\hat{\theta}^*=X_1-\text{E}_0(X_1|X_2-X_1,\cdots,X_n-X_1) \tag{2.49}$$

估计(2.49)最早由 E. J. Pitman 给出，因而也称为 Pitman 估计. 不难证明，Pitman 估计是无偏的. 事实上，因 $\hat{\theta}^*$ 是同变估计，满足(2.46)式，取 $c=-\theta$，有

$$\hat{\theta}^*(X_1-\theta,\cdots,X_n-\theta)=\hat{\theta}^*(X_1,\cdots,X_n)-\theta$$

于是

$$\begin{aligned}
\text{E}_\theta(\hat{\theta}^*-\theta)&=\text{E}_\theta\hat{\theta}^*(X_1-\theta,\cdots,X_n-\theta)\\
&=\text{E}_0\hat{\theta}^*(X_1,\cdots,X_n)\\
&=\text{E}_0(X_1)-\text{E}_0(\text{E}_0(X_1|X_2-X_1,\cdots,X_n-X_1))\\
&=0
\end{aligned}$$

Pitman 估计有一些好的性质，比如，在均方误差意义下，找不到一个一

致优于它的估计，关于 Pitman 估计的性质参见有关文献[3].

(2.49)式给出了 Pitman 估计的一般形式、进一步，我们可以算出其具体表达式. 由于在 $\theta=0$ 时，$(X_1,\cdots,X_n)\sim f(x_1,\cdots,x_n)$，作变换：$Y_1=X_1, Y_i=X_i-X_1, i=2,\cdots,n$，其 Jacobi 行列式 $|J|=1$，于是，$(Y_1,\cdots,Y_n)\sim f(y_1,y_1+y_2,\cdots,y_1+y_n)$. 在给定 $y=(y_2,\cdots,y_n)$ 下，Y_1 的条件密度函数为

$$p(y_1|y)$$
$$=f(y_1,y_1+y_2,\cdots,y_1+y_n)\Big/\int f(y_1,y_1+y_2,\cdots,y_1+y_n)\mathrm{d}y_1$$

于是

$$\mathrm{E}(Y_1|y)=\frac{\int y_1 f(y_1,y_1+y_2,\cdots,y_1+y_n)\mathrm{d}y_1}{\int f(y_1,y_1+y_2,\cdots,y_1+y_n)\mathrm{d}y_1}$$

从而

$$\hat{\theta}^*=\frac{x_1-\int y_1 f(y_1,y_1+y_2,\cdots,y_1+y_n)\mathrm{d}y_1}{\int f(y_1,y_1+y_2,\cdots,y_1+y_n)\mathrm{d}y_1}$$

进一步，作变换 $y_1=x_1-\theta$，则可将上式改写为

$$\hat{\theta}^*=\frac{\int (x_1-y_1) f(y_1,y_1+y_2,\cdots,y_1+y_n)\mathrm{d}y_1}{\int f(y_1,y_1+y_2,\cdots,y_1+y_n)\mathrm{d}y_1}$$

$$=\frac{\int \theta f(x_1-\theta,\cdots,x_n-\theta)\mathrm{d}\theta}{\int f(x_1-\theta,\cdots,x_n-\theta)\mathrm{d}\theta} \tag{2.50}$$

在利用(2.50)式求 θ 的 Pitman 估计时，必须注意积分限的确定，它是由密度函数的具体形式决定的.

例 2.39 (1)设 $X_1,\cdots,X_n$ 是来自密度函数为 $f(x-\theta)=\mathrm{e}^{-(x-\theta)}\cdot I_{(x-\theta>0)}$ 的分布的一个样本，此处 θ 是一个位置参数，由(2.50)式，θ 的 Pitman 估计为

$$\hat{\theta}=\frac{\int\theta f(x_1-\theta)\cdots f(x_n-\theta)\mathrm{d}\theta}{\int f(x_1-\theta)\cdots f(x_n-\theta)\mathrm{d}\theta}$$

$$=\frac{\int_{-\infty}^{X_{(1)}}\theta\exp\left\{n\theta-\sum_{i=1}^{n}x_i\right\}\mathrm{d}\theta}{\int_{-\infty}^{X_{(1)}}\exp\left\{n\theta-\sum_{i=1}^{n}x_i\right\}\mathrm{d}\theta}$$

$$=\frac{\frac{1}{n}X_{(1)}\exp\{nX_{(1)}\}-\frac{1}{n^2}\exp\{nX_{(1)}\}}{\frac{1}{n}\exp\{nX_{(1)}\}}$$

$$=X_{(1)}-\frac{1}{n}$$

(2)设 $X_1,\cdots,X_n$ 是来自均匀分布 $U(\theta,\theta+1)$ 的一个样本，此处 θ 是一个位置参数，$f(x-\theta)=I_{(0<x-\theta<1)}$. 由(2.50)式，$\theta$ 的 Pitman 估计为

$$\hat{\theta}=\frac{\int\theta f(x_1-\theta)\cdots f(x_n-\theta)\mathrm{d}\theta}{\int f(x_1-\theta)\cdots f(x_n-\theta)\mathrm{d}\theta}$$

$$=\frac{\int_{X_{(n)}-1}^{X_{(1)}}\theta\mathrm{d}\theta}{\int_{X_{(n)}-1}^{X_{(1)}}\mathrm{d}\theta}$$

$$=\frac{1}{2}(X_{(1)}+X_{(n)}-1)$$

(3)设 $X_1,\cdots,X_n$ 是来自 $N(\theta,\sigma^2)$ 的一个样本，$\sigma^2>0$ 已知，θ 是位置参数，$f(x-\theta)=(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(x-\theta)^2\right\}$. 由(2.50)式，$\theta$ 的 Pitman 估计为

$$\hat{\theta}=\frac{\int_{-\infty}^{\infty}\theta\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right\}\mathrm{d}\theta}{\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right\}\mathrm{d}\theta}$$

$$= \frac{\int_{-\infty}^{\infty} \theta \exp\left\{-\frac{1}{2\sigma^2}n(\overline{X}-\theta)^2\right\}\mathrm{d}\theta}{\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}n(\overline{X}-\theta)^2\right\}\mathrm{d}\theta}$$

$$= \frac{\int_{-\infty}^{\infty} t \exp\left\{-\frac{1}{2\sigma^2}nt^2\right\}\mathrm{d}t}{\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}nt^2\right\}\mathrm{d}t} + \overline{X} \quad (令\ t = \theta - \overline{X})$$

$$= \overline{X}$$

§2.7.4 尺度变换下的同变估计

下面我们考察尺度变换下的同变估计.设 $X_1,\cdots,X_n$ 是来自密度函数为 $\frac{1}{\sigma}f\left(\frac{x}{\sigma}\right)$, $x>0,\sigma>0$ 的一个样本,其中 $f(\cdot)$已知,σ 是未知的尺度参数.假如 $\hat{\sigma}=\hat{\sigma}(X_1,\cdots,X_n)$是 σ 的一个尺度变换下的同变估计,则对任意的 $\alpha>0$,应有

$$\hat{\sigma}(\alpha X_1,\cdots,\alpha X_n) = \alpha\hat{\sigma}(X_1,\cdots,X_n)$$

令 $\alpha=\frac{1}{X_n}$,记 $Y=\left(\frac{X_1}{X_n},\cdots,\frac{X_{n-1}}{X_n}\right)$,则有

$$\hat{\sigma}(X_1,\cdots,X_n) = X_n h(Y)$$

其中 $h(Y)=\hat{\sigma}(X_1/X_n,\cdots,X_{n-1}/X_n,1)$.由于$\frac{X_n}{\sigma}$和 $h(Y)$的分布均与 σ 无关,故

$$\begin{aligned} \mathrm{E}_\sigma(X_n h(Y) - \sigma)^2 &= \sigma^2\mathrm{E}_\sigma\left(\frac{X_n}{\sigma}h(Y) - 1\right)^2 \\ &= \sigma^2\mathrm{E}_1(X_n h(Y) - 1)^2 \end{aligned}$$

寻找最优同变估计等价于找 $h(\cdot)$,使 $\mathrm{E}_1(X_n h(Y)-1)^2$ 达最小.在 $\sigma=1$ 时,$(X_1,\cdots,X_n)\sim f(x_1,\cdots,x_n)$,作变换:$Y_i=\frac{X_i}{X_n}$,$i=1,\cdots,n-1$,$Y_n=X_n$,其 Jacobi 行列式 $|J|=y_n^{n-1}$,故

$$(Y_1,\cdots,Y_n) \sim f(y_1y_n,\cdots,y_{n-1}y_n,y_n)y_n^{n-1}$$

在给定 $y=(y_1,\cdots,y_{n-1})$下,Y_n 的条件分布密度函数为

$$\frac{y_n^{n-1} f(y_1 y_n, \cdots, y_{n-1} y_n, y_n)}{\int y_n^{n-1} f(y_1 y_n, \cdots, y_{n-1} y_n, y_n) \mathrm{d} y_n}$$

从而

$$\begin{aligned} & \mathrm{E}_1[(X_n h(y) - 1)^2 | y] \\ = & \frac{\int (1 - y_n h(y))^2 y_n^{n-1} f(y_1 y_n, \cdots, y_{n-1} y_n, y_n) \mathrm{d} y_n}{\int y_n^{n-1} f(y_1 y_n, \cdots, y_{n-1} y_n, y_n) \mathrm{d} y_n} \end{aligned}$$

如果 $h(y)$ 使 $\mathrm{E}_1[1 - X_n h(y))^2 | y]$ 达最小，它必然也使 $\mathrm{E}_1(1 - X_n h(Y))^2$ 达最小. 令 $r(y, h) = \int (1 - y_n h)^2 y_n^{n-1} f(y_1 y_n, \cdots, y_{n-1} y_n, y_n) \mathrm{d} y_n$，则 $r(y, h)$ 关于 h 连续，若积分与求导可交换次序，则有

$$\partial r / \partial h = 2 \int (y_n h(y) - 1) y_n^n f(y_1 y_n, \cdots, y_{n-1} y_n, y_n) \mathrm{d} y_n$$

令 $\partial r / \partial h = 0$，推出

$$h(y) = \frac{\int y_n^n f(y_1 y_n, \cdots, y_{n-1} y_n, y_n) \mathrm{d} y_n}{\int y_n^{n+1} f(y_1 y_n, \cdots, y_{n-1} y_n, y_n) \mathrm{d} y_n}$$

它使 $r(y, h)$ 达最小，亦即使 $\mathrm{E}_1[(1 - X_n h(y))^2 | y]$ 达最小. 由此给出 σ 的最优同变估计为

$$\begin{aligned} \hat{\sigma}^* &= \frac{x_n \int z^n f(y_1 z, \cdots, y_{n-1} z, z) \mathrm{d} z}{\int z^{n+1} f(y_1 z, \cdots, y_{n-1} z, z) \mathrm{d} z} \\ &= \frac{x_n \int (x_n/\sigma)^n f(x_1/\sigma, \cdots, x_{n-1}/\sigma, x_n/\sigma) x_n \sigma^{-2} \mathrm{d}\sigma}{\int (x_n/\sigma)^{n+1} f(x_1/\sigma, \cdots, x_{n-1}/\sigma, x_n/\sigma) x_n \sigma^{-2} \mathrm{d}\sigma} \quad (\text{令 } z = x_n/\sigma) \\ &= \frac{\int \sigma^{-(n+2)} f(x_1/\sigma, \cdots, x_{n-1}/\sigma, x_n/\sigma) \mathrm{d}\sigma}{\int \sigma^{-(n+3)} f(x_1/\sigma, \cdots, x_{n-1}/\sigma, x_n/\sigma) \mathrm{d}\sigma} \end{aligned} \tag{2.51}$$

例 2.40　(1)设 $X_1, \cdots, X_n$ 是来自密度函数为 $\frac{1}{\theta} f(x/\theta) = \frac{1}{\theta} \mathrm{e}^{-x/\theta}$，

$x>0$ 的指数分布的一个样本，$\theta>0$ 是尺度参数，$f(x/\theta)=\mathrm{e}^{-x/\theta}$. 记 $T=\sum_{i=1}^{n}X_i$，由(2.51)式，尺度变换下 θ 的最优同变估计为

$$\hat{\theta}=\frac{\int_0^{\infty}\theta^{-(n+2)}\mathrm{e}^{-T/\theta}\mathrm{d}\theta}{\int_0^{\infty}\theta^{-(n+3)}\mathrm{e}^{-T/\theta}\mathrm{d}\theta}=\frac{\int_0^{\infty}\sigma^{(n+2)}\mathrm{e}^{-\sigma T}\sigma^{-2}\mathrm{d}\sigma}{\int_0^{\infty}\sigma^{(n+3)}\mathrm{e}^{-\sigma T}\sigma^{-2}\mathrm{d}\sigma}$$

$$=\frac{\Gamma(n+1)T^{-(n+1)}}{\Gamma(n+2)T^{-(n+2)}}=\frac{T}{n+1}$$

(2)设 $X_1,\cdots,X_n$ 是来自均匀分布 $U(0,\theta)$ 的一个样本，$\theta>0$ 是尺度参数，$f(x/\theta)=I_{(0<x/\theta<1)}$. 由(2.51)式，$\theta$ 的最优同变估计为

$$\hat{\theta}=\frac{\int_{X_{(n)}}^{\infty}\theta^{-(n+2)}\mathrm{d}\theta}{\int_{X_{(n)}}^{\infty}\theta^{-(n+3)}\mathrm{d}\theta}=\frac{n+2}{n+1}X_{(n)}$$

例 2.41 设 $X_1,\cdots,X_n$ 是来自 $N(0,\sigma^2)$ 的一个样本，因 $X_i\in(-\infty,\infty)$，不能直接用(2.51)式，但可作变换 $T=\sum X_i^2$，于是 $T\sim Ga\left(\frac{n}{2},\frac{1}{2\sigma^2}\right)$.

(1)如果我们要估计参数 σ^2，因 T 的密度函数为

$$p(t;\sigma^2)=(2\sigma^2)^{-n/2}t^{n/2-1}\mathrm{e}^{-t/(2\sigma^2)}/\Gamma(n/2)$$

$$=\frac{1}{\sigma^2}2^{-n/2}\left(\frac{t}{\sigma^2}\right)^{n/2-1}\mathrm{e}^{-\frac{1}{2}\frac{t}{\sigma^2}}/\Gamma(n/2)$$

由(2.51)，尺度变换下 σ^2 的最优同变估计为

$$\hat{\sigma}^2=\frac{\int_0^{\infty}z^{-(1+2)}\left(\frac{t}{z}\right)^{n/2-1}\mathrm{e}^{-\frac{1}{2}\frac{t}{z}}\mathrm{d}z}{\int_0^{\infty}z^{-(1+3)}\left(\frac{t}{z}\right)^{n/2-1}\mathrm{e}^{-\frac{1}{2}\frac{t}{z}}\mathrm{d}z}$$

$$=\frac{\int_0^{\infty}z^{-(n/2+2)}\mathrm{e}^{-\frac{1}{2}\frac{t}{z}}\mathrm{d}z}{\int_0^{\infty}z^{-(n/2+3)}\mathrm{e}^{-\frac{1}{2}\frac{t}{z}}\mathrm{d}z}$$

$$=\frac{\int_0^{\infty}\theta^{n/2+2}\mathrm{e}^{-\frac{1}{2}\theta t}\theta^{-2}\mathrm{d}\theta}{\int_0^{\infty}\theta^{n/2+3}\mathrm{e}^{-\frac{1}{2}\theta t}\theta^{-2}\mathrm{d}\theta}\quad(\text{令 } z=1/\theta)$$

$$=\frac{\Gamma\left(\frac{n}{2}+1\right)\left(\frac{t}{2}\right)^{-(n/2+1)}}{\Gamma\left(\frac{n}{2}+2\right)\left(\frac{t}{2}\right)^{-(n/2+2)}}=\frac{t}{n+2}$$

(2) 如果我们要估计参数 σ，记 $S=\sqrt{T}$，则可求出 S 的密度函数为

$$\frac{1}{\sigma}\left[2^{n/2-1}\Gamma(n/2)\right]^{-1}\left(\frac{s}{\sigma}\right)^{n-1}e^{-\frac{1}{2}\left(\frac{s}{\sigma}\right)^{2}}$$

由(2.51)，尺度变换下 σ 的最优同变估计为

$$\hat{\sigma}=\frac{\int_0^{\infty}\sigma^{-3}\left(\frac{s}{\sigma}\right)^{n-1}e^{-\frac{1}{2}\left(\frac{s}{\sigma}\right)^{2}}d\sigma}{\int_0^{\infty}\sigma^{-4}\left(\frac{s}{\sigma}\right)^{n-1}e^{-\frac{1}{2}\left(\frac{s}{\sigma}\right)^{2}}d\sigma}$$

$$=\frac{\int_0^{\infty}\theta^{\frac{n+2}{2}}e^{-\frac{1}{2}s^2\theta}\,\frac{1}{2}\theta^{-3/2}d\theta}{\int_0^{\infty}\theta^{\frac{n+3}{2}}e^{-\frac{1}{2}s^2\theta}\,\frac{1}{2}\theta^{-3/2}d\theta}\quad(令\ \sigma=\theta^{-1/2})$$

$$=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n+2}{2}\right)}\frac{s}{\sqrt{2}}$$

§2.7.5　最好线性同变估计

考虑一类常见的位置——尺度分布族，其密度函数为 $\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$，$F(\cdot)$ 已知，$\mu\in\mathbf{R}$ 是位置参数，$\sigma>0$ 是尺度参数. 对此分布族，在线性变换下，寻求 μ 和 σ 的同变估计可类似前两小节的方法进行，这里我们仅限于在线性估计类中寻找最好同变估计.

定义 2.18　设 $X_1,\cdots,X_n$ 是来自某分布的一个样本. 如果参数 θ 的估计 $\hat{\theta}=\hat{\theta}(X_1,\cdots,X_n)$ 是样本的线性函数，且 $\hat{\theta}$ 是 θ 在某特定变换下的同变估计，则称 $\hat{\theta}$ 是(该变换下)θ 的线性同变估计. 假如对 θ 任一线性同变估计 $\tilde{\theta}$，还有

$$\mathrm{MSE}_\theta(\hat{\theta})\leqslant\mathrm{MSE}_\theta(\tilde{\theta}),\quad\theta\in\Theta$$

则称 $\hat{\theta}$ 是 θ 的最好线性同变估计(Best Linear Invariant Estimate)，简称 BLIE.

设 $X_1,\cdots,X_n$ 是来自密度函数为 $\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$ 的一个样本，记 $\theta=l_1\mu+l_2\sigma$，我们要估计 θ. 假如 $\hat{\theta}=l_1\hat{\mu}+l_2\hat{\sigma}=l_1\hat{\mu}(X_1,\cdots,X_n)+l_2\hat{\sigma}(X_1,\cdots,X_n)$ 是 θ 的线性变换下的线性同变估计，则对任意的 $a>0$ 和任意的 c，应有

$$\hat{\mu}(aX_1+c,\cdots,aX_n+c)=a\hat{\mu}(X_1,\cdots,X_n)+c$$
$$\hat{\sigma}(aX_1+c,\cdots,aX_n+c)=a\hat{\sigma}(X_1,\cdots,X_n)$$

从而

$$\hat{\theta}(aX_1+c,\cdots,aX_n+c)=a\hat{\theta}(X_1,\cdots,X_n)+l_1c \tag{2.52}$$

我们在 2.5 节的最后部分已经介绍过 μ 和 σ 的 BLUE，令 μ^* 和 σ^* 分别表示由(2.41)式给出的 μ 和 σ 的 BLUE，其协差阵由(2.42)式给出，记为

$$\operatorname{Var}\begin{pmatrix}\mu^*\\ \sigma^*\end{pmatrix}=\sigma^2\begin{pmatrix}A & B\\ B & C\end{pmatrix}$$

在(2.52)式中取 $a=1/\sigma^*$，$c=-\mu^*/\sigma^*$，则有

$$\begin{aligned}&\hat{\theta}(X_1,\cdots,X_n)/\sigma^*-l_1\mu^*/\sigma^*\\ &=\hat{\theta}((X_1-\mu^*)/\sigma^*,\cdots,(X_n-\mu^*)/\sigma^*)\end{aligned}$$

记 $h=h(\mu^*,\sigma^*,X)=\hat{\theta}((X_1-\mu^*)/\sigma^*,\cdots,(X_n-\mu^*)/\sigma^*)$，得到 $\hat{\theta}$ 的一般形式：

$$\hat{\theta}_h=l_1\mu^*+h\sigma^*$$

其均方误差为

$$\begin{aligned}\mathrm{E}(\hat{\theta}_h-\theta)^2&=\mathrm{E}(l_1\mu^*+h\sigma^*-l_1\mu-l_2\sigma)^2\\ &=\mathrm{E}[(\sigma^*-\sigma)h+l_1(\mu^*-\mu)+(h-l_2)\sigma]^2\\ &=h^2\operatorname{Var}(\sigma^*)+l_1^2\operatorname{Var}(\mu^*)+(h-l_2)^2\sigma^2+2hl_1\operatorname{Cov}(\mu^*,\sigma^*)\\ &=\sigma^2[Ch^2+Al_1^2+(h-l_2)^2+2hl_1B]\end{aligned}$$

将上式对 h 求导并令导数等于 0 可知，当 $h=(l_2-l_1B)/(1+C)$ 时 $\mathrm{MSE}(\hat{\theta}_h)$ 达最小，从而给出 θ 的 BLIE 为

$$\hat{\theta} = l_1\mu^* + \frac{l_2 - l_1 B}{1 + C}\sigma^* \tag{2.53}$$

这是 θ 的有偏估计.其均方误差比 θ 的 BLUE$\theta^* = l_1\mu^* + l_2\sigma^*$ 的均方误差要小.

特别,若分别取 $l_1=0, l_2=1$ 和 $l_1=1, l_2=0$,则分别给出 σ 和 μ 的 BLIE 为

$$\hat{\sigma} = \frac{1}{1+C}\sigma^*, \quad \hat{\mu} = \mu^* - \frac{B}{1+C}\sigma^* \tag{2.54}$$

例 2.42　设 $X_{(1)} < \cdots < X_{(r)}$ 是来自 I 型极值分布 $F(x) = 1 - \exp\{-\exp[(x-\mu)/\sigma]\}$ 的容量为 n 的样本的前 r 个次序观测值,在例 2.37 中,我们给出了 σ 和 μ 的 BLUE(见(2.43)和(2.44)),其协差阵由(2.42)式给出.易见,在 n 和 r 给定下,A, B, C 均与 σ, μ 无关,由(2.54)式可写出 σ 和 μ 的 BLIE 为

$$\hat{\sigma} = \sum_{j=1}^{r} C_I(n,r,j)X_{(j)}, \quad \hat{\mu} = \sum_{j=1}^{r} D_I(n,r,j)X_{(j)} \tag{2.55}$$

其中,$C_I(n,r,j) = \frac{1}{1+C}C(n,r,j)$ 称为 σ 的最好线性同变估计系数,$D_I(n,r,j) = \frac{1}{1+C}D(n,r,j) - \frac{B}{1+C}C(n,r,j)$ 称为 μ 的最好线性同变估计系数,它们的值可在文献[6]中表一查到.譬如,利用表 2.2 中的数据,在文献[6]中表一中查到诸 $C_I(n,r,j)$ 和 $D_I(n,r,j)$,由(2.55)式可计算出 σ 和 μ 的 BLIE 为 $\hat{\sigma} = 1.2122$,$\hat{\mu} = 4.4114$(见表 2.3).

表 2.3　极值分布中参数的 BLIE 计算表($n=8, r=5$)

j	$X_{(j)}$	$C_I(n,r,j)$	$C(n,r,j)X_{(j)}$	$D_I(n,r,j)$	$D(n,r,j)X_{(j)}$
1	0.9163	−0.1802	−0.1651	−0.0627	−0.0575
2	2.0149	−0.1765	−0.3556	−0.0322	−0.0649
3	2.8332	−0.1496	−0.4238	0.0128	0.0363
4	3.7955	−0.1016	−0.3856	0.0724	0.2748
5	4.1821	0.6079	2.5423	1.0097	4.2227
求和			1.2122		4.4114

参考文献

1 Sorenson H W. Parameter Estimation: Principles and Problems. Marcel Dekker, 1980

2 Kiefer J C. Introduction to Statistical Inference. Springer-Verlag, 1989

3 成平,陈希孺,陈桂景,吴传义. 参数估计. 上海:上海科学技术出版社,1985

4 陈希孺. 数理统计引论. 北京:科学出版社,1984

5 陈希孺,方兆本,李国英,陶波. 非参数统计. 上海:上海科学技术出版社,1989

6 可靠性试验用表. 国防工业出版社,1979

7 Rao C R. Linear Statistical Inference and Its Applications(Second Edition). John Wiley and Sons, 1975

8 Cox D R, Hinkley D V. Theroretical Statistics. Chapman and Hall, 1974

9 王松桂. 线性模型的理论及其应用. 合肥:安徽教育出版社,1987

习 题 二

2.1 设 X_1, X_2 独立同分布,其共同的密度函数为

$$p(x;\theta) = 3x^2/\theta^3, \quad 0 < x < \theta, \quad \theta > 0$$

(1) 证明 $T_1 = \frac{2}{3}(x_1 + x_2)$ 和 $T_2 = \frac{7}{6}\max(x_1, x_2)$ 都是 θ 的无偏估计;

(2) 计算 T_1 和 T_2 的均方误差并进行比较;

(3) 证明:在均方误差意义下,在形如 $T_c = c\max(x_1, x_2)$ 的估计中,$T_{8/7}$ 最优.

2.2 设 X_1, X_2 独立同分布,其共同密度函数为

$$p(x;\theta)=k\theta^k x^{-(k+1)},\quad x>\theta,\quad \theta>0,\quad k>2 \text{ 已知}.$$

(1) 证明 $T_1=\frac{k-1}{2k}(x_1+x_2)$ 和 $T_2=\frac{2k-1}{2k}\min(x_1,x_2)$ 都是 θ 的无偏估计；

(2) 计算 T_1 和 T_2 的均方误差并进行比较；

(3) 证明：在均方误差意义下，在形如 $T_c=c\min(x_1,x_2)$ 的估计中，$c=\frac{2k-2}{2k-1}$ 时最优；

(4) 如果 $1<k\leqslant 2$，则 T_1 的方差为无穷大而 T_2 的方差有限；如果 $k=1$，你用什么估计？

2.3　设 $\theta\in(a,b)$，$T(X)$ 是 θ 的无偏估计，令

$$S(x)=\begin{cases}T(x), & a\leqslant T(x)\leqslant b\\ a, & T(x)<a\\ b, & T(x)>b\end{cases}$$

证明：$\mathrm{E}(S(X)-\theta)^2\leqslant \mathrm{E}(T(X)-\theta)^2$.

2.4　(Jackknife)设 $X_1,\cdots,X_n$ 是来自某分布的一个样本，$T_n=T_n(X_1,\cdots,X_n)$ 是参数 $g(\theta)$ 的估计，并设

$$\mathrm{E}T_n-g(\theta)=\sum_{r=1}^{\infty}a_r/r!$$

a_r 可以是 θ 的函数. 记 $T_{n-1}^{(i)}$ 表示用剔除 X_i 后容量为 $n-1$ 的剩余样本以同样方法得到的 θ 的估计，

$$\overline{T}_{n-1}=\frac{1}{n}\sum_{i=1}^{n}T_{n-1}^{(i)}$$

$$T'_n=nT_n-(n-1)\overline{T}_{n-1}$$

证明：$\mathrm{E}T'_n-g(\theta)=O(n^{-2})$. 这表明 T'_n 的偏差比 T_n 低了一阶，T'_n 称为由 T_n 得到的一步 Jackknife 估计. 类似地，可给出二步，三步 Jackknife估计.

2.5　设 $X_1,\cdots,X_n$ 是来自 $b(1,\theta)$，$\theta\in(0,1)$ 的样本，对 $g(\theta)=\theta^2$，一个直观估计是

$$T_n(X)=\frac{1}{n}\sum_{i=1}^{n}X_i^2$$

试用 Jackknife 方法对 T_n 进行减偏.

2.6 设 $\hat{\theta}_n$ 是 θ 的估计量,证明:若 $n\to\infty$ 时,$E\hat{\theta}_n\to\theta$,$\mathrm{Var}(\hat{\theta}_n)\to 0$,则 $\hat{\theta}_n$ 是 θ 的相合估计.

2.7 $X_1,\cdots,X_n$ 独立同分布,$EX_1=\mu$,$\mathrm{Var}(X_1)<\infty$,证明:$\hat{\mu}=\frac{2}{n(n+1)}\sum_{i=1}^{n} iX_i$ 是 μ 的相合估计.

2.8 设 $X_1,\cdots,X_n$ 是来自 $U(0,n\lambda)$ 的一个样本,

(1) 证明 $n\lambda-X_{(n)}$ 的分布收敛于 $\mathrm{Exp}(\lambda)$;

(2) 用(1)中结果给出 λ 的一个相合估计.

2.9 设 $(X_1,Y_1),\cdots,(X_n,Y_n)$ 为独立同分布的二元正态变量,$EX_1=EY_1=0$,$\mathrm{Var}X_1=\mathrm{Var}Y_1=1$,$\mathrm{Cov}(X_1,Y_1)=\rho$. 记

$$S_{xx}=\frac{1}{n}\sum X_i^2,\quad S_{xy}=\frac{1}{n}\sum X_iY_i,\quad S_{yy}=\frac{1}{n}\sum Y_i^2$$

(1) 证明 $\sqrt{n}(S_{xx}-1,S_{xy}-\rho,S_{yy}-1)\xrightarrow{L}N_3(0,\boldsymbol{\Sigma})$,其中

$$\boldsymbol{\Sigma}=\begin{pmatrix}2 & 2\rho^2 & \rho^2\\ 2\rho & 1+\rho^2 & 2\rho\\ 2\rho^2 & 2\rho & 2\end{pmatrix}$$

(2) 令 $r=S_{xy}/\sqrt{S_{xx}S_{yy}}$,求 $\sqrt{n}(r-\rho)$ 的渐近分布.

2.10 设 $X_1,\cdots,X_n$ 是来自 $b(1,\theta),\theta\in\Theta$ 的一个样本.

(1) 证明:若 Θ 的元素超过 n 个,则 $S(x)=\sum_{i=1}^{n}x_i$ 是 θ 的完备充分统计量;

(2) 证明:机会比率 $\theta/(1-\theta)$ 是不可估的.

2.11 设 $X_1,\cdots,X_n$ 是来自均值为 μ,方差为 σ^2 的分布的一个样本,μ,σ^2 均未知,考虑 μ 的线性估计类

$$\mathscr{L}_\theta=\{T(x):T(x)=\sum_{i=1}^{n}c_ix_i\}$$

(1) 证明:$T(x)$ 为 θ 的无偏估计的充要条件是 $\sum c_i=1$;

(2) 证明:$\overline{X}$ 在线性无偏估计类中方差一致达到最小.

2.12 考虑如下离散型的均匀分布 $P_\theta(X=i)=\frac{1}{\theta},i=1,\cdots,\theta$. 记

$\mathscr{P}=\{P_\theta,\theta=1,2,\cdots\}$.

(1)试证 $\mathscr{P}$ 是完备的；

(2)试求 θ 的 UMVUE$\hat{\theta}$；

(3)固定某正整数 k，令 $\mathscr{P}_k=\mathscr{P}-\{P_k\}$，试证 $\mathscr{P}_k$ 是不完备的；

(4)试证对 $\mathscr{P}_k$，$\hat{\theta}$ 不是 θ 的 UMVUE.

2.13　考虑幂级数分布的参数估计问题. 设 $X_1,\cdots,X_n$ 是来自

$$Pr(X=x)=\frac{a_x}{f(\theta)}\theta^x,\quad x=c,c+1,\cdots,\infty \tag{2.56}$$

的一个样本，令 $T=\sum_{i=1}^n x_i$.

(1)证明：T 是 θ 的完备充分统计量，且 T 的分布仍具(2.56)形式

$$Pr(T=t)=\frac{b_t}{f^n(\theta)}\theta^t,\quad t=nc,n+1,\cdots,\infty$$

(2)证明 θ^r 的 UMVUE 是

$$\hat{\theta}^r=\begin{cases}0, & t<r\\ b_{t-r}/b_t, & t\geqslant r\end{cases}$$

(3)证明：若 X 的取值范围有限，则(1)，(2)中的结论不再正确.

2.14　设 $X_1,\cdots,X_n$ 是来自 $N(\theta,1)$的一个样本，令 $g(\theta)=Pr(X_1\leqslant 0)$，试求 $g(\theta)$的 UMVUE.

2.15　设 $X_1,\cdots,X_n$ 是来自对数正态分布 $LN(\mu,\sigma^2)$的一个样本，试求均值 EX_1 的 UMVUE.

2.16　(1)设 $X\sim p(x;\theta)=f(x)/h(\theta)$，$\theta\leqslant x\leqslant b$. θ 是未知参数，$h(\theta)>0$，$f(x)>0$. 证明，若 $T(X)$是 $g(\theta)$的无偏估计，则

$$-T(X)=\{g(X)h'(X)+g'(X)h(X)\}/f(X)$$

(2)设 $X_1,\cdots,X_n$ 是来自(1)中分布的一个样本，则 $g(\theta)$的唯一的 UMVUE 为

$$T(X)=g(X_{(1)})-\frac{g'(X_{(1)})}{n}\frac{h(X_{(1)})}{f(X_{(1)})}$$

类似地可讨论 $p(x;\theta)=f(x)/h(\theta)$，$a\leqslant x\leqslant\theta$ 的情形.

2.17　设 $X_0,X_1,\cdots,X_n$ 服从 $n+1$ 元正态分布，且 $EX_i=0$，$i=0,1,\cdots,n$，

$$\mathrm{Cov}(X_i,X_j)=\begin{cases}\sigma^2, & i=j\\ \rho\sigma^2, & i\neq j\end{cases}\quad \sigma^2>0,\ -1<\rho<1.$$

(1)证明:$X_0,X_1,\cdots,X_n$ 的联合分布是指数族,并求(σ^2,ρ)的完备充分统计量;

(2)求 σ^2 和 $\rho\sigma^2$ 的 UMVUE.

(提示:令 $\boldsymbol{J}$ 表示全部元素由 1 组成的矩阵,验证$(a\boldsymbol{I}+b\boldsymbol{J})^{-1}=c\boldsymbol{I}+d\boldsymbol{J}$,$a\neq0$.)

2.18 (1) 设 $T(x)$是 $g(\theta)$的无偏估计,$\mathrm{Var}(T(X))<\infty$. 证明:$T(x)$为 $g(\theta)$的 UMVUE 的充要条件是,对任一 0 的无偏估计$\varphi(x)$,若 $\mathrm{Var}(\varphi(X))<\infty$,则 $\mathrm{Cov}(\varphi(X),T(X))=0$;

(2)试用(1)的结论证明,若 $X_1,\cdots,X_n$ 是来自 $N(\mu,\sigma^2)$的一个样本,则 $\bar{X}$ 和 s_{n-1}^2分别是 μ 和 σ^2 的 UMVUE.(提示:对 0 的任一无偏估计 T_0,对 $\mathrm{E}T_0=0$ 的运算式求导.)

2.19 设 X 的概率分布为

$$P_\theta(X=k)=\begin{cases}\theta, & k=-1\\ \theta^k(1-\theta)^2, & k=0,1,\cdots\end{cases}\quad \theta\in(0,1)$$

(1)找出 0 的无偏估计类 U_0;

(2)证明 θ 有无偏估计但其 UMVUE 不存在. 这表明,该分布族没有完备充分统计量;

(3)令 $g(\theta)=(1-\theta)^2$,证明 $g(\theta)$的 UMVUE 存在.

2.20 设 $T(x)$是 $g(\theta)$的 UMVUE,对 $g(\theta)$的任一无偏估计 $\hat{g}(x)$,证明:若 $\mathrm{Var}(\hat{g}(X))<\infty$,则 $\mathrm{Cov}(T,\hat{g})\geqslant0$. 且 $\hat{g}$ 也为 $g(\theta)$的 UMVUE 的充要条件是 $T(x)$与 $\hat{g}(x)$的相关系数等于 1.

2.21 证明:若 $T_1(X)$,$T_2(X)$分别是 $g_1(\theta)$,$g_2(\theta)$的 UMVUE,则 $T_1(X)+T_2(X)$是 $g_1(\theta)+g_2(\theta)$的 UMVUE.

2.22 设 $T(X)$是 $g(\theta)$的 UMVUE,$T_1,T_2\in U_g$,且 $\mathrm{Var}(T_i)=k_i\mathrm{Var}(T)$,$i=1,2$,$k_i$ 可与 θ 有关. 记

$$\rho=\frac{\mathrm{Cov}(T_1,T_2)}{(\mathrm{Var}(T_1)\mathrm{Var}(T_2))^{1/2}},\quad \alpha_i=\frac{1}{k_i},i=1,2$$

证明:

$$(\alpha_1\alpha_2)^{1/2} - [(1-\alpha_1)(1-\alpha_2)]^{1/2} \leqslant \rho$$
$$\leqslant (\alpha_1\alpha_2)^{1/2} + [(1-\alpha_1)(1-\alpha_2)]^{1/2}$$

特别，若 $S(X)\in U_g$，且 $\mathrm{Var}(S)=k\mathrm{Var}(T)$，则 T 与 S 的相关系数 $\rho(S,T)=k^{-1/2}$.（提示：令 $T_a=aT_1+(1-a)T_2\in U_g$，由 $\mathrm{Var}(T_a)\geqslant \mathrm{Var}(T)$ 给出 $a^2k_1+(1-a)^2k_2+2a(1-a)\rho(k_1k_2)^{1/2}\geqslant 1$，且至多有一个 a 使该式中等号成立.）

2.23　(1)设 $T(X)$ 是 $g(\theta)$ 的 UMVUE，$U(X)\in U_g$，定义 $V(X)$ 满足 $aU+(1-a)V=T$，a 是常数，$0<a<1$. 证明

$$\mathrm{Var}(V)-\mathrm{Var}(U)\begin{cases}<0, & a<1/2\\ =0, & a=1/2\\ >0, & a>1/2\end{cases}$$

(2)设 $X_1,\cdots,X_n$ 是来自 $N(0,\theta)$ 的样本，定义

$$V=\frac{(n-2)\sum_{i=1}^{n}X_i^2+(n\overline{X})^2}{n(n-1)}$$

$$U=\frac{\sum(X_i-\overline{X})^2}{n-1}$$

证明：U 与 V 有相同的均值和方差.

2.24　对 Poisson 分布 $P(\theta)$，

(1) 求 $I(\theta)$；

(2) 求 $I\left(\frac{1}{\theta}\right)$；

(3) 找一个函数 $g(\cdot)$，使 $g(\theta)$ 的 Fisher 信息与 θ 无关.

2.25　设 $X_1,\cdots,X_n$ 独立同分布. X_1 的取值有四种可能，其概率分别为 $p_1=1-\theta$，$p_2=\theta-\theta^2$，$p_3=\theta^2-\theta^3$，$p_4=\theta^3$. 以 N_j 记 $X_1,\cdots,X_n$ 中出现各种可能结果的次数，$N_1+N_2+N_3+N_4=n$.

(1) 确定 a_1,a_2,a_3,a_4，使 $T=\sum_{j=1}^{4}a_jN_j$ 为 θ 的无偏估计；

(2) 将 $\mathrm{Var}(T)$ 与 θ 的无偏估计方差的 C-R 下界比较.

2.26　设 $\{p(x;\theta),\theta\in(a,b)\}$ 为一概率密度函数族，假定

(1) $\int x^2p(x;\theta)\mathrm{d}x \triangleq \alpha_2(\theta)<\infty$；

(2) $\int p(x;\theta)\mathrm{d}x = 1$, $\int xp(x;\theta)\mathrm{d}x \triangleq \alpha_1(\theta)$ 可在积分号下求导;

(3) $I(\theta)=\mathrm{E}(\partial\ln p(x;\theta)/\partial\theta)^2<\infty$,则

$$I(\theta) \geqslant \frac{(\partial\alpha_1(\theta)/\partial\theta)^2}{\sigma^2(\theta)}$$

其中 $\sigma^2(\theta)=\alpha_2(\theta)-(\alpha_1(\theta))^2$.

2.27 设 X 具概率密度函数 $p(x;\theta)=\theta p_1(x)+(1-\theta)p_2(x)$,$0<\theta<1$,$p_1(x)$,$p_2(x)$是两个完全已知的概率密度函数,其支撑不依赖于 θ. 设 $X_1,\cdots,X_n$ 是来自该分布的一个样本,证明:θ 的无偏估计的 C-R 下界为

$$\theta(1-\theta)/(n(1-J)) \tag{2.57}$$

其中 $J=\int_{-\infty}^{\infty}\frac{p_1(x)p_2(x)}{p(x;\theta)}\mathrm{d}x$. 如果 $p_1(x)$,$p_2(x)$的支撑不相交,则 $J=0$,(2.75)即为二项分布中 θ 的无偏估计的 C-R 下界. 二点分布是 $p_1(x)=x$,$p_2(x)=1-x$,$x=0,1$ 时的特例.

2.28 (1) 设 $T(X)$是 $g(\theta)$的无偏估计,并设 Cramer-Rao 正则条件满足,证明:$T(X)$是 $g(\theta)$的有效无偏估计的充要条件是存在 $a(\theta)$,使得

$$\frac{\partial l}{\partial\theta}=a(\theta)(T(x)-g(\theta))$$

其中 $l=l(\theta;x)$表示对数似然函数;

(2) 在(1)中,我们指出,若存在 $g(\theta)$的无偏估计 $T(X)$,使 $T(X)-g(\theta)$是$\frac{\partial l}{\partial\theta}=\frac{1}{L}\frac{\partial L}{\partial\theta}$的线性函数($L$ 是似然函数),则 $T(X)$是 $g(\theta)$的有效无偏估计. 在多数情况下,找不到满足这样条件的 $T(X)$,然而,有时可能存在 $g(\theta)$的无偏估计 $T(X)$满足 $T(X)-g(\theta)$是$\frac{1}{L}\frac{\partial L}{\partial\theta},\frac{1}{L}\frac{\partial^2 L}{\partial\theta^2},\cdots$的线性函数,它也是最优的无偏估计,达到下面介绍的 Bhattacharyya 下界,它改进了 C-R 下界.

假设 Cramer-Rao 正则条件满足,记 $L^{(r)}=\frac{\partial^r L}{\partial\theta^r}$,$g^{(r)}=\frac{\partial^r g(\theta)}{\partial\theta^r}$,$r=1,\cdots,k$. 假定 $I_{rs}=\mathrm{E}\left(\frac{L^{(r)}}{L}\frac{L^{(s)}}{L}\right)\neq 0$,$r,s=1,\cdots,k$,$T(X)$是 $g(\theta)$的无偏

估计，证明：

$$\mathrm{Var}(T) \geqslant \sum_{r=1}^{k}\sum_{s=1}^{k} g^{(s)} I_{rs}^{-1} g^{(r)}$$

且等号成立的充要条件是

$$T(x) - g(\theta) = \sum_{r=1}^{k}\left(\sum_{s=1}^{k} g^{(s)} I_{rs}^{-1}\right)\frac{L^{(r)}}{L}$$

此处 $\sum_{r=1}^{k}\sum_{s=1}^{k} g^{(s)} I_{rs}^{-1} g^{(r)}$ 称为 $g(\theta)$ 的无偏估计的 Bhattacharyya 下界；

(3) 设 $X_1,\cdots,X_n$ 是来自 $N(\theta,1)$ 的样本，记 $T(X)=\overline{X}^2-\frac{1}{n}$，证明：$T(X)$ 是 θ^2 的无偏估计，且 $T(X)-\theta^2$ 是 $\frac{L^{(1)}}{L}$ 和 $\frac{L^{(2)}}{L}$ 的线性函数，从而 $T(X)$ 是 θ^2 的 UMVUE 且方差达到 $k=2$ 时的 Bhattacharyya 下界.

2.29 设 $X_1,\cdots,X_n$ 是来自 $b(1,\theta)$ 的样本，$0<\theta<1$.

(1) 求 $\theta(1-\theta)$ 的 UMVUE $T(X)$；

(2) 证明：$T(X)$ 的方差达不到 $\theta(1-\theta)$ 的无偏估计的 C-R 下界，但达到 $k=2$ 时的 Bhattacharyya 下界.

2.30 设 $X_1,\cdots,X_n$ 是来自二项分布 $b(k,p)$ 的样本，k,p 未知.

(1) 证明：$n=1$ 时参数不可估；

(2) 设 $n\geqslant 2$，求 k,p 的矩估计；

(3) 求(2)中估计的渐近分布.

2.31 设 (X_i,Y_i)，$i=1,\cdots,n$ 为独立同分布变量，$Pr(X_1>0)=1$，$EX_1^2<\infty$，$EY_1^2<\infty$. 定义 $\theta=EY_1/EX_1$，证明

$$\sqrt{n}\left(\frac{\overline{Y}}{\overline{X}}-\theta\right)\xrightarrow{L} N(0,V)$$

并求 V.

2.32 设 $X_1,\cdots,X_n$ 是来自密度函数为 $p(x;\theta)=\theta(1+\theta)x^{\theta-1}(1-x)$，$0<x<1$，$\theta>0$ 的一个样本.

(1) 求 θ 的矩估计 $\hat{\theta}_n$；

(2) 求 $\sqrt{n}(\hat{\theta}_n-E\hat{\theta}_n)$ 的渐近分布；

(3) $\hat{\theta}_n$ 是否是渐近有效的？

2.33 $X_1,\cdots,X_n$ 为独立同分布变量，$Pr(X=k)=p_k(\theta)$，$k=1,\cdots,K$. 以 N_k 表示 $X_i=k$ 的个数，设 $p_k(\theta)$ 对所有 θ 可微，导数不为 0.

(1) 固定 k，求由 $p_k(\hat{\theta}_k)=N_k/n$ 得到的估计 $\hat{\theta}_k$ 的渐近分布；

(2) $\hat{\theta}_k$ 是否是相合估计？给一个 $\hat{\theta}_k$ 的渐近方差的相合估计.

2.34 设 $X_1,\cdots,X_n$ 是来自 $N(0,\sigma^2)$ 的一个样本.

(1) 利用二阶矩构造 σ^2 的矩估计，进而给出 σ 的矩估计 $\hat{\sigma}_1$；

(2) 证明 $E_\sigma|X_1|=\sigma\sqrt{\frac{2}{\pi}}$. 由此给出 σ 的另一矩估计 $\hat{\sigma}_2$；

(3) 比较 $\hat{\sigma}_1$ 和 $\hat{\sigma}_2$ 的 MSE，解释比较结果.

2.35 设 $X_1,\cdots,X_n$ 为独立同分布变量，$0<\theta<1$，

$$Pr(X_1=-1)=\frac{1-\theta}{2},\quad Pr(X_1=0)=1/2$$

$$Pr(X_1=1)=\frac{\theta}{2}$$

(1) 求 θ 的 MLE $\hat{\theta}_1$ 并问 $\hat{\theta}_1$ 是否是无偏的；

(2) 求 θ 的矩估计 $\hat{\theta}_2$；

(3) 计算 θ 的无偏估计的方差的 C-R 下界.

2.36 设 $X_1,\cdots,X_n$ 是来自 $U(0,\theta)$，$\theta>0$ 的一个样本，则 θ 的 MLE 是 $\hat{\theta}_1=X_{(n)}$，UMVUE 是 $\hat{\theta}_2=\frac{n+1}{n}X_{(n)}$. 试证：存在一个估计 $\hat{\theta}(x)$，使

$$\mathrm{MSE}_\theta(\hat{\theta})<\min(\mathrm{MSE}_\theta(\hat{\theta}_1),\mathrm{MSE}_\theta(\hat{\theta}_2))$$

2.37 设 $X_1,\cdots,X_n$ 为独立同分布变量，$X_1\sim p(x;\theta)=\theta x^{\theta-1}$，$0<x<1$，$\theta>0$. 令 $\eta=1/\theta$.

(1) 求 η 的 MLE $\hat{\eta}$；

(2) 计算 $E\hat{\eta}$ 和 $\mathrm{Var}(\hat{\eta})$；

(3) 求 η 的 UMVUE.

2.38 设 $X_1,\cdots,X_n$ 是来自 $N(\mu,\sigma^2)$ 的一个样本. 记 $\theta=(\mu,\sigma^2)$，固定 c，令 $q_c(\theta)$ 表示 c 下方的总体的比例.

(1) 试求 $q_c(\theta)$ 的 MLE；

(2) 试证 $q_c(\theta)$的 UMVUE 为

$$T(x)=\begin{cases}0, & kV\leqslant -1\\ G\left(\frac{1}{2}+\frac{1}{2}kV\right), & kV\in(-1,1)\\ 1, & kV\geqslant 1\end{cases}$$

其中

$$k=\sqrt{n}/(n-1)$$

$$V=(c-\overline{X})/\hat{\sigma}\quad\left(\overline{X}=\frac{1}{n}\sum X_i,\hat{\sigma}^2=\frac{1}{n-1}\sum(X_i-\overline{X})^2\right)$$

$G(\cdot)$表示 Beta 分布 $Be\left(\frac{1}{2}(n-2),\frac{1}{2}(n-2)\right)$的分布函数.

2.39　设 $X_1,\cdots,X_n$ 是来自 $U(\theta_1,\theta_2)$的一个样本.

(1) 证明$(X_{(1)},X_{(n)})$是该分布族的完备充分统计量;

(2) 求 θ_1,θ_2 的一致最小方差无偏估计;

(3) 证明:若 $\theta_2=2\theta_1$,则$(X_{(1)},X_{(n)})$不再是完备的;

(4) 对 $\theta_2=2\theta_1$,求 θ_1 的 MLE,对之作修偏,记为 $\hat{\theta}_1$;

(5) 令 $T=(n+1)(2X_{(n)}+X_{(1)})/(5n+4)$,则 T 是 θ_1 的无偏估计,且 $\mathrm{Var}(T)\leqslant\mathrm{Var}(\hat{\theta}_1)$.

2.40　设 $X_1,\cdots,X_n$ 是来自$(k\theta,(k+1)\theta)$上均匀分布的样本,$k>0$ 已知.

(1) 求 θ 的 MLE,记为 $\hat{\theta}$;

(2) 证明 $\mathrm{Var}(X_{(1)})=\mathrm{Var}(X_{(n)})$;

(3) 令 $\tilde{\theta}=X_{(1)}/k$,考虑形如 $T_a=a\hat{\theta}+(1-a)\tilde{\theta}$ 的估计,求使 $\mathrm{Var}(T_a)$达最小的 a,记为 θ^*,计算 $\mathrm{Var}(\theta^*)/\mathrm{Var}(\hat{\theta})$.

2.41　设 $X_1,\cdots,X_n$ 是来自$\{p(x;\mu,\sigma)=\frac{1}{2\sigma}\exp\{-|x-\mu|/\sigma\},\mu\in R,\sigma\in\mathbf{R}^+\}$的一个样本.

(1) 求 μ,σ 的矩估计;

(2) 求 μ,σ 的极大似然估计.

2.42　设 $X_1,\cdots,X_n$ 是来自下列 Weibull 分布的一个样本,

$$F(x)=1-\exp\{-x^m/\eta\},\quad x>0$$

若其中 m 已知，

(1) 试求 η 的 MLE$\hat{\eta}$；

(2) 试求 $\hat{\eta}$ 的渐近分布.

2.43 证明：对正态分布 $N(\mu,\sigma^2)$，若只有一个观测值，则 μ,σ^2 的极大似然估计不存在.

2.44 $X_1,\cdots,X_n$ 是来自 $N(\theta,a\theta^2)$ 的一个样本，$\theta>0$ 未知，$a>0$ 已知.

(1) 计算 θ 的 MLE 并求其渐近分布；

(2) 对什么 a 值，$\overline{X}$ 的渐近效超过 0.9?

2.45 $X_i\sim Exp(\alpha)$，$i=1,\cdots,n$，$Y_j\sim Exp(\alpha\theta)$，$j=1,\cdots,n$，合样本独立. 求 θ 的 MLE$\hat{\theta}$ 并将它的渐近方差与 α 已知时 θ 的 MLE 的渐近方差进行比较.

2.46 考虑幂级数分布的极大似然估计. $X_1,\cdots,X_n$ 为来自(2.56)的一个样本，$\overline{X}=\dfrac{1}{n}\sum X_i$.

(1) 证明 $\mathrm{E}_\theta X\triangleq\mu(\theta)=\theta f'(\theta)/f(\theta)$；

(2) 证明 θ 的 MLE 是 $\overline{X}=\mu(\theta)$ 的根. 这里的似然方程与矩法方程一致；

(3) 求单个观测值的 Fisher 信息 $I_1(\theta)$；

(4) 试对 Poisson 分布，二项分布，截尾 Poisson 分布分别写出上述似然方程的显式表示式.

2.47 设 $X_i\sim N(\mu,\omega_i\sigma^2)$，$i=1,\cdots,n$，$\omega_i$ 已知，诸 X_i 独立.

(1) 试求 μ 的 BLUE$\hat{\mu}$；

(2) 问 $\hat{\mu}$ 是否是 μ 的有效无偏估计；

(3) 设 σ^2 已知，求位移变换下 μ 的最优同变估计.

2.48 设 $X_i\sim N(\mu,\sigma^2)$，$i=1,\cdots,m$，$Y_i\sim N(c\mu,\sigma^2)$，$i=1,\cdots,n$，合样本独立，c 已知.

(1) 试求 μ 的 BLUE$\hat{\mu}$；

(2) $\hat{\mu}$ 是否是 μ 的有效无偏估计.

(3) 设 σ^2 已知，求位移变换下 μ 的最优同变估计.

2.49　设 $X_1,\cdots,X_n$ 独立同分布，X_i 的密度函数为 kx^{k-1}/θ^k，$0<x<\theta$，$k>0$ 已知，$\theta>0$ 未知. 试在尺度变换下求 θ 的最优同变估计.

2.50　设 $X_1,\cdots,X_n$ 独立同分布，其共同的密度函数为 $4\theta^4x^{-5}$，$x>\theta>0$，θ 是未知参数，试在尺度变换下求 θ 的最优同变估计.

2.51　设 $X_1,\cdots,X_n$ 是来自密度函数为 $\frac{1}{2\sigma}\exp\{-|x|/\sigma\}$ 的双指数分布的一个样本，$\sigma>0$ 是未知参数，试在尺度变换下求 σ 的最优同变估计.

第三章　假设检验

§3.1　基本概念

常用的U检验，t检验和F检验都具有足够的直观性和可信感，本章将对这些检验的优良性作深入的研究．那么，一个检验"好"的标准是什么？对于给定的"好坏"标准，是否存在最优的检验方法？如果存在，又怎样求得？这些问题的深入研究就要涉及假设检验理论的一些基本问题．为此，先引进一些基本概念．

§3.1.1　假设

定义 3.1　设$(\mathscr{X},\mathscr{B},\mathscr{P})$为一统计结构，则$\mathscr{P}$的非空子集称为假设．在参数分布族$\mathscr{P}=\{P_\theta;\theta\in\Theta\}$时，$\Theta$的非空子集称为假设．

设样本空间$\mathscr{X}$中的样本点x是我们的观测值，除了参数估计，统计推断中的又一类重要问题是，检验观测值x与给定的假设是否存在矛盾．这类问题称为假设检验问题．

在一个假设检验问题中常涉及两个假设．所要检验的假设称为原假设，记为H_0．与H_0不相容的假设称为备择假设，记为H_1．关于统计结构$(\mathscr{X},\mathscr{B},\mathscr{P})$的原假设和备择假设分别记为

$$H_0: P\in\mathscr{P}_0 \text{ 对 } H_1: P\in\mathscr{P}_1$$

这里$\mathscr{P}_0$和$\mathscr{P}_1$是$\mathscr{P}$的两个互不相交的非空子集．在参数分布族$\mathscr{P}=\{P_\theta:\theta\in\Theta\}$场合，原假设和备择假设分别记为

$$H_0:\theta\in\Theta_0 \text{ 对 } H_1:\theta\in\Theta_1$$

这里Θ_0和Θ_1是Θ的两个互不相交的非空子集．给定H_0和H_1就等于给定一个检验问题，记为检验问题(H_0,H_1)．而(Θ_0,Θ_1)称为参数假设检验问题，其它称为非参数假设检验问题．

§3.1.2　检验、拒绝域与检验统计量

定义 3.2　在检验问题(H_0, H_1)中，所谓检验法则(简称检验法，或检验)，就是设法把样本空间划分为互不相交的两个可测集：

$$\mathscr{X} = W + \overline{W}$$

并作如下规定：

当观测值 $x \in W$ 时，就拒绝原假设 H_0，认为备择假设 H_1 成立；

当观测值 $x \notin W$(即 $x \in \overline{W}$)时，就不拒绝原假设 H_0.

这里的 W 称为检验的拒绝域.

这样一来，选定了检验法，就是确定了拒绝域；反之，选定了拒绝域，也就确定了检验法.

为了确定拒绝域，往往首先由问题的直观背景出发，寻找一个统计量，使得在原假设 H_0 成立时和在备择假设 H_1 成立时，该统计量的值有差异. 从而使得我们能够根据这个统计量的值的大小选定拒绝域. 我们称这个能从样本空间中划分出拒绝域的统计量为检验统计量.

例 3.1　电话交换台单位时间内接到的呼唤次数服从泊松分布 $P(\lambda)$，$\lambda > 0$. λ 为单位时间内接到的平均呼唤次数. 为了考察该交换台在单位时间内的平均呼唤次数是否不超过 1，可考虑建立如下两个假设：

$$\text{原假设 } H_0 : \lambda \leqslant 1 \quad \text{对} \quad \text{备择假设 } H_1 : \lambda > 1 \qquad (3.1)$$

设 $x = (x_1, \cdots, x_n)$ 是该电话交换台的 n 次记录，即它是来自 Poisson 分布 $\mathscr{P} = P(\lambda)$ 的样本 $X = (X_1, \cdots, X_n)$ 的一个观测值. 取检验统计量 $T = \sum_{i=1}^{n} X_i$. 它是 λ 的充分，完备的统计量. 一般来说，在原假设 H_0 成立时检验统计量的值较小，而在备择假设 H_1 成立时检验统计量的值较大. 这里有一个临界值 c，在 $T \geqslant c$ 时认为样本数据 x 与原假设相矛盾，因而拒绝原假设；而在 $T < c$ 时就不拒绝原假设. 该检验的拒绝域可用检验统计量表示如下：

$$W = \{x : \sum_{i=1}^{n} x_i \geqslant c\}$$

这个检验是直观和可信的.如何确定 c 值的问题与检验犯两类错误的概率有关.

§3.1.3 两类错误

在进行检验时,由于样本的随机性,我们可能作出正确的判断,也可能作出错误的判断,正确的判断是原假设 H_0 成立时接受 H_0,或原假设 H_0 不成立时拒绝 H_0. 而错误的判断是原假设 H_0 成立时但被拒绝,或原假设 H_0 不成立时但被接受.为了对检验法的好坏给出一个合理的评选标准,需要考察一个检验法可能犯错误的概率.

当原假设 H_0 成立时,样本观测值却落在拒绝域中 W,从而拒绝了原假设.这种错误称为第一类错误.犯第一类错误的概率为

$$\alpha = P(X \in W), \quad P \in \mathscr{P}_0$$

在参数统计结构,犯第一类错误的概率为

$$\alpha(\theta) = P_\theta(X \in W), \quad \theta \in \Theta_0$$

另一类错误是,当原假设 H_0 不成立时,样本观测值却没有落在拒绝域 W 中,从而没有拒绝原假设.这种错误称为第二类错误.犯第二类错误的概率为

$$\beta = P(X \notin W) = 1 - P(X \in W), \quad P \in \mathscr{P}_1$$

在参数统计结构,犯第二类错误的概率为

$$\beta(\theta) = P_\theta(X \notin W) = 1 - P_\theta(X \in W), \quad \theta \in \Theta_1$$

譬如例 3.1 中的检验犯第一类错误的概率为

$$\alpha(\lambda) = \sum_{k=c}^{\infty} \frac{(n\lambda)^k}{k!} \cdot \exp\{-n\lambda\}, \quad \lambda \leqslant 1 \tag{3.2}$$

而检验犯第二类错误的概率为

$$\begin{aligned}\beta(\lambda) &= \sum_{k=0}^{c-1} \frac{(n\lambda)^k}{k!} \cdot \exp\{-n\lambda\} \\ &= 1 - \sum_{k=c}^{\infty} \frac{(n\lambda)^k}{k!} \cdot \exp\{-n\lambda\}, \quad \lambda > 1\end{aligned} \tag{3.3}$$

这两类错误的概率都可以通过 $P_\theta(X \in W)$ 来表示.

§3.1.4　势函数

定义 3.3　称样本观察值落在拒绝域的概率为检验的势函数，记为

$$g(\theta)=P_\theta(X\in W),\quad \theta\in\Theta$$

在 $\theta\in\Theta_0$ 时，$g(\theta)=\alpha(\theta)$，$g(\theta)$是检验犯第一类错误的概率. 在 $\theta\in\Theta_1$时，$g(\theta)=1-\beta(\theta)$，$1-g(\theta)$是检验犯第二类错误的概率.

在例 3.1 中，检验的势函数为

$$g(\lambda)=\sum_{k=c}^{\infty}\frac{(n\lambda)^k}{k!}\cdot\exp\{-n\lambda\}\tag{3.4}$$

它是 λ 的严增函数. 在 $n=10$ 时，临界值分别取为 $c=16$ 和 $c=14$ 的两个检验的势函数的具体曲线形式如图 3.1 所示.

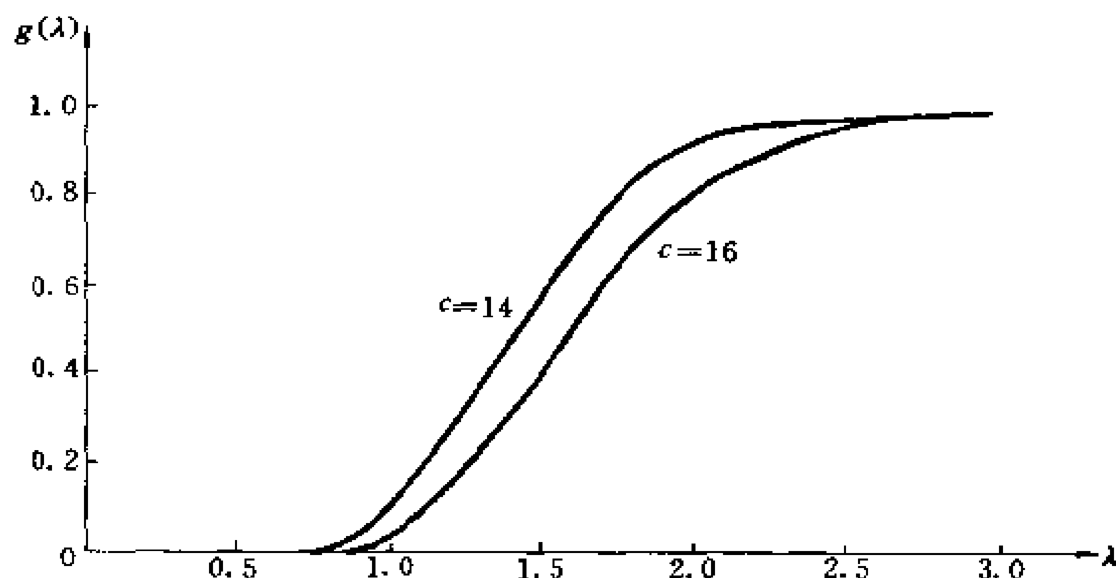


图 3.1　Poisson 分布均值检验的势函数的图像

§3.1.5　检验的水平

在例 3.1 中，从(3.2)和(3.3)两式可以看出，在样本容量 n 固定时，要减少犯第一类错误的概率，必须增加 c，从而导致增大犯第二类错误的概率；反之，若要减少犯第二类错误的概率，必须减少 c，从而使

犯第一类错误的概率增加.换句话说,当样本容量 n 固定时,不可能使得犯这两类错误的概率都减少.这一现象在一般的检验问题中也都出现.

基于这种情况,需要采取某种妥协方案.Neyman 和 Pearson 的假设检验理论的基本思想,就是使得犯第一类错误的概率限制在某一个范围内,然后寻找使犯第二类错误的概率尽可能小的检验.在这种思想指导下,寻找一个好的检验法,就是对选定的一个较小的数 $\alpha(0<\alpha<1)$,在满足

$$g(\theta)=P_\theta(X\in W)\leqslant\alpha,\quad \theta\in\Theta_0$$

的检验中,寻找这样的检验,使得在 $\theta\in\Theta_1$ 时,$g(\theta)$ 尽可能的大.

在 $\theta\in\Theta_0$ 时,$g(\theta)\leqslant\alpha$ 的检验称为水平为 α 的检验,记为 $(\alpha,\Theta_0,\Theta_1)$ 检验.α 是事先选定的,由于 α 的大小反映了检验犯第一类错误的概率的大小,所以常取一个较小的数.例如取 0.1,0.05,0.01 等作为 α 的值.α 的选取也依赖于我们关于假设的先验信息,依赖于我们对犯第二类错误的要求.比如,根据以往的经验,非常相信原假设是真的,而犯第二类错误又不会造成大的影响或后果,此时 α 就可以取得小一些.又比如,第二类错误带来的影响较大,需要严格控制犯第二类错误的概率,此时 α 可以选得适当大一些.

在例 3.1 中,由于检验的势函数 $g(\lambda)$(见 3.4 式)是 λ 的严增函数,给定 α 后,为使犯第一类错误的概率不超过 α,即在 $\lambda\leqslant1$ 时,$g(\lambda)\leqslant\alpha$,只需

$$g(1)=\sum_{k=c}^{\infty}\frac{n^k}{k!}\cdot\exp\{-n\}\leqslant\alpha \tag{3.5}$$

$n=10$,c 取不同的值时,(3.5)式左边的值如下表所示.

表 3.1 Poisson 分布均值检验犯第一类错误的概率

c	$g(l)$	c	$g(l)$
14	0.135 537	17	0.027 043
15	0.083 460	18	0.014 279
16	0.048 742	19	0.007 188

在 $n=10$ 时,若取 $\alpha=0.05$,由于

$$\sum_{k=16}^{\infty}\frac{10^k}{k!}\cdot\exp\{-10\}=0.048\,740<0.05 \tag{3.6}$$

$$\sum_{k=15}^{\infty}\frac{10^k}{k!}\cdot\exp\{-10\}=0.083\,458>0.05 \tag{3.7}$$

根据 Neyman 和 Pearson 的假设检验理论的基本思想,取 $c=16$. 所以在 $X_1,\cdots,X_{10}$是来自泊松分布总体 $\mathscr{P}=(P(\lambda),\lambda>0)$的一个样本时,关于检验问题(3.1)的水平 $\alpha=0.05$ 的拒绝域为

$$W=\left\{x:\sum_{i=1}^{10}x_i\geqslant 16\right\}$$

在上述的叙述中,对拒绝原假设,给以特别的注意. 在原假设为真时,样本观测值落入拒绝域是一个小概率事件. 根据实际推断原理,在一次观测中,小概率事件是几乎不可能发生的. 所以拒绝原假设的理由是充足的. 由于检验犯第二类错误的概率并没有受到限制,仅要求它尽可能的小,所以在备择假设为真时,样本观测值没有落入拒绝域可能不是一个小概率事件. 因此接受原假设的理由可能是不充足的. 较为科学的说法是不拒绝原假设. 把"不拒绝原假设"说成"接受原假设",这是为了防止在诸如"除了拒绝与接受之外,还有第三种可能"之类的哲理性的讨论,从而使实际工作者能较快地熟悉和使用假设检验方法.

对给定的水平 α,人们总是力图构造这样的检验,使得检验犯第一类错误的概率

$$g(\theta)=P_\theta(X\in W)\leqslant\alpha,\quad \theta\in\Theta_0$$

并且至少存在一个 $\theta\in\Theta_0$,使得

$$g(\theta)=P_\theta(X\in W)=\alpha$$

这意味着水平被"足量"地使用了,以满足尽可能减少检验犯第二类错误的概率的要求. 在例 3.1 中,$n=10$ 时,所求得的水平 $\alpha=0.05$ 的检验,水平并没有被"足量"的使用. 在原假设 $H_0:\lambda\leqslant 1$ 成立时,

$$\alpha(\lambda)=\sum_{i=16}^{\infty}\frac{(10\lambda)^k}{k!}\cdot\exp\{-10\lambda\}$$

$$\leqslant \sum_{k=16}^{\infty} \frac{10^k}{k!} \cdot \exp\{-10\} = 0.048\,740$$

为了能构造出水平 $\alpha=0.05$ 被“足量”地使用的检验，需要引进随机化检验.

§3.1.6 检验函数和随机化检验

定义函数

$$\phi(x) = \begin{cases} 1, & x \in W \\ 0, & x \notin W \end{cases}$$

它是拒绝域 W 的示性函数，仅取 0 和 1 两个值. 反之，如果一个函数 $\phi(x)$ 仅取 0 和 1 两个值，则 $W=\{x : \phi(x)=1\}$ 可作为拒绝域. 于是，给出了这样的一个函数，就等于给了一个检验法. 这样的函数 $\phi(x)$ 称为检验函数，简称检验. 它是仅取 0 和 1 两个值的统计量. 检验的势函数可表示为

$$g(\theta) = P_\theta(X \in W) = \mathrm{E}_\theta(\phi(X)) \tag{3.8}$$

为了能构造出水平被“足量”地使用的检验，需要拓广检验函数的概念.

定义 3.4 设 $\phi(x)$ 为定义在 $\mathscr{X}$ 上的可测函数，满足条件 $0 \leqslant \phi(x) \leqslant 1$，则称 $\phi(x)$ 为检验函数，简称检验. 在 $\phi(x)$ 仅取 0,1 两个值时，则为非随机化检验. 否则，为随机化检验. 其势函数为 $g(\theta)=\mathrm{E}_\theta\phi(X)$.

非随机化检验的拒绝域 $W=\{x : \phi(x)=1\}$. 在随机化检验时，假设样本观测值为 x，则我们以概率 $\phi(x)$ 拒绝原假设；以概率 $1-\phi(x)$ 接受原假设. 在 $\theta \in \Theta_0$ 时，$g(\theta)$ 是检验犯第一类错误的概率，而在 $\theta \in \Theta_1$ 时，$1-g(\theta)$ 是检验犯第二类错误的概率.

在例 3.1 中，$n=10$，$\alpha=0.05$ 时，考虑到检验的直观和可信，取非随机化检验函数

$$\phi(x) = \begin{cases} 1, & T \geqslant 16 \\ r, & T = 15 \\ 0, & T \leqslant 14 \end{cases}$$

其中 $T = \sum_{i=1}^{n} x_i$. 在原假设 $H_0 : \lambda \leqslant 1$ 成立时，该检验犯第一类错误的概率为

$$\alpha(\lambda)=\mathrm{E}_\lambda\phi(X)=\sum_{k=16}^{\infty}\frac{(10\lambda)^k}{k!}\cdot\exp\{-10\lambda\}+$$

$$r\cdot\frac{(10\lambda)^{15}}{15!}\cdot\exp\{-10\lambda\}$$

$$\leqslant\sum_{k=16}^{\infty}\frac{10^k}{k!}\cdot\exp\{-10\}+r\cdot\frac{10^{15}}{15!}\cdot$$

$$\exp\{-10\}=0.048\,740+r\cdot 0.034\,718$$

为使得水平 $\alpha=0.05$ 被"足量"地使用，应有

$$0.048\,740+r\cdot 0.034\,718=0.05,r=\frac{0.05-0.048\,740}{0.034\,718}=0.036$$

这个非随机化检验的实施过程如下：若样本观察值 $x_1,\cdots,x_{10}$满足条件 $\sum_{i=1}^{10}x_i\geqslant16$，则拒绝原假设；若 $\sum_{i=1}^{10}x_i\leqslant14$，则接受原假设；若 $\sum_{i=1}^{10}x_i=15$，先做一个成功概率为 0.036 的贝努里试验，如果试验结果为成功，则拒绝原假设，反之则接受原假设.

在下面两节，这个随机化检验的优良性将被说明.

§3.1.7　充分性原则

定义 3.5　设 $\phi_1(x)$和 $\phi_2(x)$都是某检验问题(H_0,H_1)的检验函数，如果它们的势函数相同，即

$$\mathrm{E}_\theta[\phi_1(X)]=\mathrm{E}_\theta[\phi_2(X)],\quad \theta\in\Theta$$

则称检验函数 $\phi_1(x)$和 $\phi_2(x)$等价.

这个定义意味着，检验函数的统计性质完全取决于其势函数. 势函数相同，就认为这两个检验函数等价.

定理 3.1　设 $X=(X_1,\cdots,X_n)$是来自分布族 $\mathscr{P}=(P_\theta:\theta\in\Theta)$的样本. $T(X)$是关于 θ 的充分统计量. 则对任意一个检验函数 $\phi(x)$，存在另一个只依赖于 $T(x)$的检验函数，它与 $\phi(x)$相互等价.

证明：因为 $T(X)$是关于 θ 的充分统计量，所以条件期望 $\varphi(t)=\mathrm{E}[\phi(X)|T=t]$与 θ 无关，又因为 $0\leqslant\phi(x)\leqslant1$，所以 $0\leqslant\varphi(t)\leqslant1$. 因而 $\varphi(t)$也是检验函数. 由于

$$\mathrm{E}_\theta\varphi(T(X))=\mathrm{E}_\theta\{\mathrm{E}[\phi(X)|T]\}=\mathrm{E}_\theta(\phi(X))$$

所以 $\varphi(t)$ 和 $\phi(x)$ 具有相同的势函数，相互等价. 证毕.

这个定理告诉我们，当 θ 的充分统计量存在时，关于 θ 的任何假设检验问题，其优良检验，只要在充分统计量构成的检验函数中去寻找就可以了. 这就是假设检验中所谓的"充分性原则".

§3.2 Neyman-Pearson 基本引理

如果一个假设只含有一个元素，则称该假设为简单假设，否则称为复合假设. 在例 3.1 中，原假设 $H_0: \lambda \leqslant 1$ 和备择假设 $H_1: \lambda > 1$ 都是复合假设. 如果我们考虑的检验问题是，原假设 $H_0: \lambda = 1$ 对备择假设 $H_1: \lambda > 1$，那么这个原假设就是简单假设.

现在我们来考虑简单原假设对简单备择假设的检验问题. 这不仅是因为这样的检验问题特别简单，已经完满地解决，而且也是我们解决许多更一般的检验问题的基础.

设 $(\mathscr{X}, \mathscr{B}, \{P_\theta : \theta \in \Theta\})$ 为一参数统计结构，考虑检验问题：

$$\text{简单原假设 } H_0: \theta = \theta_0 \text{ 对简单备择假设 } H_1: \theta = \theta_1 (\theta_0 \neq \theta_1) \tag{3.9}$$

这时，比较任意两个水平为 α 的检验 $\phi_1(x)$ 和 $\phi_2(x)$ 的好坏是容易的. 在 $\mathrm{E}_{\theta_1}[\phi_1(X)] \geqslant \mathrm{E}_{\theta_1}[\phi_2(X)]$ 时，我们说 $\phi_1(x)$ 不比 $\phi_2(x)$ 差. 由之，很自然地有下面的定义.

定义 3.6 在检验问题 (Θ_0, Θ_1) 中，设 $\phi(x)$ 是水平为 α 的检验. 如果对任意一个水平为 α 的检验 $\phi_1(x)$，都有

$$\mathrm{E}_{\theta_1}[\phi(X)] \geqslant \mathrm{E}_{\theta_1}[\phi_1(X)]$$

则称检验 $\phi(x)$ 是水平为 α 的最优势检验，记为 MPT(Most Powerful Test).

下面的著名引理证明了，在简单原假设对简单备择假设的检验问题中，MPT 一定存在，并且可以具体构造出 MPT 的检验函数.

定理 3.2 (Neyman-Pearson 基本引理，简称 N-P 基本引理) 设 P_{θ_0} 和 P_{θ_1} 是可测空间 $(\mathscr{X}, \mathscr{B})$ 上的两个不同的概率测度，关于某个 σ 有

限的测度 μ,有

$$p(x;\theta_0)=\frac{\mathrm{d}P_{\theta_0}}{\mathrm{d}\mu},\quad p(x;\theta_1)=\frac{\mathrm{d}P_{\theta_1}}{\mathrm{d}\mu}$$

则在检验问题(3.9)中,

(1)对给定的 $\alpha(0<\alpha<1)$存在一个检验函数 $\phi(x)$及常数 $k\geqslant0$,使得

$$\mathrm{E}_{\theta_0}\phi(X)=\alpha \tag{3.10}$$

$$\phi(x)=\begin{cases}1, & p(x;\theta_1)>k\cdot p(x;\theta_0)\\ 0, & p(x;\theta_1)<k\cdot p(x;\theta_0)\end{cases} \tag{3.11}$$

(2)由(3.10)和(3.11)两式确定的检验函数 $\phi(x)$是水平为 α 的 MPT. 反之,如果 $\phi(x)$是水平为 α 的 MPT,则一定存在常数 $k\geqslant0$,使得 $\phi(x)$满足(3.11)式(a.e.$[\mu]$).

证明:1° 首先证明(1),存在形如(3.11)的检验函数,且使(3.10)式成立.

对于任一实数 λ,令

$$G(\lambda)=P_{\theta_0}\{p(X;\theta_1)>\lambda\cdot p(X;\theta_0)\}$$

由于这个概率是在 P_{θ_0}下计算的,所以只要在集合$\{x:p(x;\theta_0)>0\}$上考虑不等式 $p(x;\theta_1)>\lambda\cdot p(x;\theta_0)$.

$G(\lambda)$是非负随机变量$\dfrac{p(X;\theta_1)}{p(X;\theta_0)}>\lambda$的概率,故 $1-G(\lambda)$是随机变量$\dfrac{p(X;\theta_1)}{p(X;\theta_0)}$的分布函数. 所以 $G(\lambda)$是一个非增、右连续函数,且

$$G(+\infty)=0,\quad G(0-0)=1$$

$$G(\lambda-0)-G(\lambda)=P_{\theta_0}\left\{\frac{p(X;\theta_1)}{p(X;\theta_0)}=\lambda\right\}$$

给定 $\alpha\in(0,1)$后,有且只有下列两种情况:

Ⅰ. 存在 $\lambda_0\geqslant0$,使得 $G(\lambda_0)=\alpha$. 定义

$$\phi(x)=\begin{cases}1, p(x;\theta_1)>\lambda_0\cdot p(x;\theta_0)\\ 0, p(x;\theta_1)\leqslant\lambda_0\cdot p(x;\theta_0)\end{cases}$$

则

$$\mathrm{E}_{\theta_0}\phi(X) = P_{\theta_0}\{p(X;\theta_1) > \lambda_0 \cdot p(X;\theta_0)\} = G(\lambda_0) = \alpha$$

Ⅱ. 存在 $\lambda_0 \geqslant 0$，使得 $G(\lambda_0) < \alpha \leqslant G(\lambda_0 - 0)$. 定义

$$\phi(x) = \begin{cases} 1, & p(x;\theta_1) > \lambda_0 \cdot p(x;\theta_0) \\ \dfrac{\alpha - G(\lambda_0)}{G(\lambda_0 - 0) - G(\lambda_0)}, & p(x;\theta_1) = \lambda_0 \cdot p(x;\theta_0) \\ 0, & p(x;\theta_1) < \lambda_0 \cdot p(x;\theta_0) \end{cases}$$

则

$$\begin{aligned} \mathrm{E}_{\theta_0}\phi(X) = & P_{\theta_0}\{p(X;\theta_1) > \lambda_0 \cdot p(X;\theta_0)\} + \\ & \frac{\alpha - G(\lambda_0)}{G(\lambda_0 - 0) - G(\lambda_0)} \cdot P_{\theta_0}\{p(X;\theta_1) \\ & = \lambda_0 \cdot p(X;\theta_0)\} = G(\lambda_0) + [\alpha - G(\lambda_0)] = \alpha \end{aligned}$$

在这两种情况下，λ_0 就可以被取作(3.11)式中的非负常数 k. $\phi(x)$ 使得(3.10)式成立，这说明 $\phi(x)$ 是水平为 α 的检验函数. 它是水平 α 被"足量"地使用的检验函数.

2° 其次证明，由公式(3.10)，(3.11)得到的检验函数 $\phi(x)$ 是 MPT.

设 $\phi^*(x)$ 是其它任意一个水平为 α 的检验函数，即 $\mathrm{E}_{\theta_0}\phi^*(X) \leqslant \alpha$. 由于 $\phi(x)$ 满足(3.11)式，所以

$$[\phi(x) - \phi^*(x)] \cdot [p(x;\theta_1) - k \cdot p(x;\theta_0)] \geqslant 0$$

因此

$$\int [\phi(x) - \phi^*(x)] \cdot [p(x;\theta_1) - k \cdot p(x;\theta_0)] \mathrm{d}\mu(x) \geqslant 0$$

从而

$$\begin{aligned} \mathrm{E}_{\theta_1}\phi(X) - \mathrm{E}_{\theta_1}\phi^*(X) & \geqslant k \cdot [\mathrm{E}_{\theta_0}\phi(X) - \mathrm{E}_{\theta_0}\phi^*(X)] \\ & = k \cdot [\alpha - \mathrm{E}_{\theta_0}\phi^*(X)] \geqslant 0 \end{aligned}$$

所以 $\phi(x)$ 是 MPT.

3° 最后证明：若 $\phi^*(x)$ 是水平为 α 的 MPT，则一定存在非负常数 k，使得 $\phi^*(x)$ 满足(3.11)式(a.e.$[\mu]$).

设 $\phi(x)$ 是满足(3.10)，(3.11)两式的检验函数. 由 2° 知，$\phi(x)$ 是水平为 α 的 MPT. 由于 $\phi^*(x)$ 和 $\phi(x)$ 都是水平为 α 的 MPT，则

$E_{\theta_1}\phi^*(X) - E_{\theta_1}\phi(X)$. 由于 $\phi(x)$ 满足(3.11)式，所以

$$[\phi(x) - \phi^*(x)] \cdot [p(x;\theta_1) - k \cdot p(x;\theta_0)] \geqslant 0 \tag{3.12}$$

因为 $E_{\theta_1}\phi^*(X) = E_{\theta_1}\phi(X)$, $E_{\theta_0}\phi^*(X) \leqslant \alpha = E_{\theta_0}\phi(X)$，所以

$$\int [\phi(x) - \phi^*(x)] \cdot [p(x;\theta_1) - k \cdot p(x;\theta_0)] d\mu(x)$$

$$= E_{\theta_1}\phi(X) - E_{\theta_1}\phi^*(X) - k \cdot [E_{\theta_0}\phi(X) - E_{\theta_0}\phi^*(X)] \leqslant 0 \tag{3.13}$$

比较(3.12)和(3.13)两个不等式，可得

$$[\phi(x) - \phi^*(x)] \cdot [p(x;\theta_1) - k \cdot p(x;\theta_0)] = 0, (\text{a.s.}[\mu])$$

故在集合 $\{x: p(x;\theta_1) - k \cdot p(x;\theta_0) \neq 0\}$ 上，有 $\phi^*(x) = \phi(x)$ (a.s. $[\mu]$). 这证明了 $\phi^*(x)$ 满足(3.11)式(a.e. $[\mu]$). 证毕.

注 1. 满足(3.11)式的检验函数 $\phi(x)$ 通常称为似然比检验函数(或称为概率比检验函数). 在集合 $\{x: p(x;\theta_0) > 0$ 或 $p(x;\theta_1) > 0\}$ 上，定义似然检验比函数

$$\lambda(x) = \frac{p(x;\theta_1)}{p(x;\theta_0)}$$

在 $\lambda(x)$ 比较大的时候，$p(x;\theta_1)$ 比较大，所以原假设 $H_0: \theta = \theta_0$ 为真时观察到样本点 x 的可能性比备择假设 $H_1: \theta = \theta_1$ 为真时观察到样本点 x 的可能性小. 而在 $\lambda(x)$ 比较小的时候，$p(x;\theta_0)$ 比较大，所以备择假设 $H_1: \theta = \theta_1$ 为真时观察到样本点 x 的可能性比原假设 $H_0: \theta = \theta_0$ 为真时观察到样本点 x 的可能性小. 由此很自然地，我们在 $\lambda(x)$ 比较大的时候，拒绝原假设，认为 $\theta = \theta_1$ 成立；而在 $\lambda(x)$ 比较小的时候，不拒绝原假设，认为 $\theta = \theta_0$ 成立. N-P 基本引理告诉我们，在简单原假设对简单备择假设的检验问题中，MPT 是似然比检验；反此，似然比检验也是 MPT.

注 2. 在似然比 $\lambda(x)$ 具有连续分布函数时，MPT 检验函数可取为非随机化的形式

$$\phi(x) = \begin{cases} 1, & \lambda(x) \geqslant k \\ 0, & \lambda(x) < k \end{cases}$$

其中 k 由(3.10)式，即 $E_{\theta_0}\phi(x) = P_{\theta_0}\{\lambda(x) \geqslant k\} = \alpha$ 确定. 在 $\lambda(x)$ 为离

散随机变量时，至多在集合$\{x : \lambda(x)=k\}$上实施随机化. MPT 检验函数可取为

$$\phi(x)=\begin{cases}1, & \lambda(x)>k\\ r, & \lambda(x)=k\\ 0, & \lambda(x)<k\end{cases}$$

其中k与r由(3.10)式，即由以下两式确定

$$P_{\theta_0}\{\lambda(X)\geqslant k\}\geqslant\alpha\geqslant P_{\theta_0}\{\lambda(X)>k\}=\alpha_1$$

$$r=\frac{\alpha-\alpha_1}{P_{\theta_0}\{\lambda(X)=k\}}$$

有时，在集合$\{x : \lambda(x)=k\}$上可以实施非随机化(见下面的例 3.3).

注 3. 在非参数统计结构$(\mathscr{X},\mathscr{B},\mathscr{P})$中，考虑简单原假设$H_0 : P=P_0$对简单备假设择$H_1 : P=P_1$的检验问题. 设$P_0$和$P_1$是可测空间$(\mathscr{X},\mathscr{B})$上的两个不同的概率测度，且关于某个$\sigma$有限的测度$\mu$，有

$$p_0(x)=\frac{\mathrm{d}P_0}{\mathrm{d}\mu},\quad p_1(x)=\frac{\mathrm{d}P_1}{\mathrm{d}\mu}$$

定义似然比$\lambda(x)=\dfrac{p_1(x)}{p_0(x)}$. N-P 基本引理仍然适用.

例 3.2 设$X=(X_1,\cdots,X_n)$是来自正态分布族$\{N(\mu,1):-\infty<\mu<\infty\}$的样本，考虑如下的检验问题：

原假设$H_0 : \mu=0$ 对 备择假设$H_1 : \mu=\mu_1(\mu_1>0)$

取水平为$\alpha(0<\alpha<1)$. 构造似然比统计量

$$\lambda(x)=\frac{\prod_{i=1}^{n}p(x_i;\mu_1)}{\prod_{i=1}^{n}p(x_i;0)}=\exp\{n\mu_1\bar{x}-\frac{1}{2}n\mu_1^2\}$$

MPT 的拒绝域具有形式

$$W=\{x : \lambda(x)\geqslant k\}=\{x : \bar{x}\geqslant c\}$$

当H_0成立时，$\overline{X}\sim N\left(0,\dfrac{1}{n}\right)$. 所以对于给定的水平$\alpha$，$c=\dfrac{U_{1-\alpha}}{\sqrt{n}}$. MPT 检验仅与水平$\alpha$有关，而与$\mu_1$的具体数值无关. 只要求$\mu_1>0$就行了.

如果把此例中的备择假设改为$H_1 : \mu=\mu_1(\mu_1<0)$，则 MPT 检验

的拒绝域为

$$W = \left\{ x : x \leqslant \frac{U_\alpha}{\sqrt{n}} \right\}$$

同样，该检验仅与水平 α 有关，而与 μ_1 的具体数值无关，只要求 $\mu_1<0$ 就行了.

若正态分布族为 $\{N(\mu,\sigma_0^2): -\infty<\mu<\infty\}$，$H_0: \mu=\mu_0$，$H_1: \mu=\mu_1(\mu_1>\mu_0)$，则可以通过变换 $y_i=\dfrac{x_i-\mu_0}{\sigma_0}$，将这个假设检验问题归结为上面所述的情况. 其拒绝域为

$$W = \left\{ \frac{\overline{x}-\mu_0}{\sigma_0} \geqslant \frac{U_{1-\alpha}}{\sqrt{n}} \right\}$$

例 3.3　设 $X=(X_1,\cdots,X_n)$ 是来自均匀分布族 $\{R(0,\theta): \theta>0\}$ 的样本，考虑如下的检验问题：

原假设 $H_0: \theta=1$　对　备择假设 $H_1: \theta=\theta_1(\theta_1>1)$

取水平为 $\alpha(0<\alpha<1)$. 构造似然比统计量

$$\lambda(x) = \frac{\prod_{i=1}^{n} p(x_i;\theta_1)}{\prod_{i=1}^{n} p(x_i;1)} = \begin{cases} \theta_1^{-n}, & 0<x_{(n)}<1 \\ \infty, & 1 \leqslant x_{(n)} < \theta_1 \end{cases}$$

其中 $x_{(n)}=\max\{x_1,\cdots,x_n\}$. 在原假设 H_0 成立时，$\lambda(x)$ 的分布是退化分布. 令 $G(\lambda)=P(\lambda(X)>\lambda|\theta=1)$，则

$$0 = G(\theta_1^{-n}) < \alpha < G(\theta_1^{-n}-0) = 1$$

由 N-P 基本引理知，取 $k=\theta_1^{-n}$. 则 MPT 为

$$\phi(x) = 1 \qquad 1 \leqslant x_{(n)} < \theta_1$$

而在 $0<x_{(n)}<1$ 时，$\phi(x)$ 的值由条件 $\mathrm{E}[\phi(X)|\theta=1]=\alpha$ 而定. 例如，取随机化检验

$$\phi(x) = \alpha \qquad 0 < x_{(n)} < 1$$

也可取非随机化检验

$$\phi(x) = \begin{cases} 0, & 0 < x_{(n)} < c \\ 1, & c \leqslant x_{(n)} < 1 \end{cases}$$

由于在原假设 $H_0: \theta=1$ 成立时,$T=X_{(n)}$ 的密度函数为 $p(t)=nt^{n-1}(0<t<1)$,故由 $\mathrm{E}[\phi(X)|\theta=1]=\alpha$ 推得,$c=\sqrt[n]{1-\alpha}$. 显然,后一个非随机化检验

$$\phi(x)=\begin{cases}0, & 0<x_{(n)}<\sqrt[n]{1-\alpha}\\ 1, & \sqrt[n]{1-\alpha}\leqslant x_{(n)}<\theta_1\end{cases} \tag{3.14}$$

较为合理.

如果把此例中的备择假设改为 $H_1: \theta=\theta_1(\theta_1<1)$,其 MPT 如何?习题 3.3 将讨论这个问题.

如果原假设和备择假设分别为 $H_0: \theta=\theta_0$ 和 $H_1: \theta=\theta_1(\theta_1>\theta_0)$,则可以通过变换 $y_i=\dfrac{x_i}{\theta_0}$,将这个假设检验问题归结为上面所述的情况. 其非随机化检验为

$$\phi(x)=\begin{cases}0, & 0<y_{(n)}<\sqrt[n]{1-\alpha}\\ 1, & \sqrt[n]{1-\alpha}\leqslant y_{(n)}<\theta_1\end{cases}$$

其中 $y_{(n)}=\max\{y_1,\cdots,y_n\}=\dfrac{x_{(n)}}{\theta_0}$.

例 3.4 设 $X=(X_1,\cdots,X_n)$ 是来自泊松分布族 $\{P(\lambda): \lambda>0\}$ 的样本,考虑如下的检验问题:

原假设 $H_0: \lambda=1$ 对备择假设 $H_1: \lambda=\lambda_1(\lambda_1>1)$

取水平为 $\alpha(0<\alpha<1)$. 构造似然比统计量

$$\lambda(x)=\frac{\prod_{i=1}^{n}p(x_i;\lambda_1)}{\prod_{i=1}^{n}p(x_i;1)}=(\lambda_1)^{\sum x_i}\cdot\exp\{-(\lambda_1-1)\}$$

$\lambda(x)$ 关于 $T=\sum_{i=1}^{n}x_i$ 严格单调上升. 根据 N P 基本引理,水平 $\alpha=0.05$ 的 MPT 检验函数满足条件

$$\phi(x)=\begin{cases}1, & T>k\\ 0, & T<k\end{cases}$$

在 $n=10$ 时,若原假设 $H_0: \lambda=1$ 成立,$T\sim P(10)$. 则由(3.6)和(3.7)两式知,取 $k=15$. 则检验函数为

$$\phi(x)=\begin{cases}1, & T>15\\ r, & T=15\\ 0, & T<15\end{cases}$$

同例 3.1，由水平 $\alpha=0.05$ 算得 $r=0.036$. 这个检验与 λ_1 的具体数值无关，只要求 $\lambda_1>0$ 就行了.

上述三例的 MPT 检验函数都仅依赖于充分统计量，而这正如定理 3.1 的充分性原则告诉我们的. 事实上，可以基于充分统计量，而不是样本构造似然比统计量. 这样求得的检验函数仍如上面所述的.

关于 MPT 还有一个性质，它在下一节的定理的证明中将要用到.

定理 3.3　设 $\phi(x)$ 是定理 3.2 中水平为 α 的最优势检验函数，则 $E_{\theta_1}\phi(X)\geqslant\alpha$. 又设 $0<\alpha<1$. 则除了 $p(x;\theta_1)=p(x;\theta_0)$ a.e. $[\mu]$ 外，必有 $E_{\theta_1}\phi(X)>\alpha$.

证明：令 $\phi_1(x)\equiv\alpha$. 显然，$\phi_1(x)$ 是水平为 α 的检验. 因为 $\phi(x)$ 是水平为 α 的最优势检验函数，故 $E_{\theta_1}\phi(X)\geqslant E_{\theta_1}\phi_1(X)=\alpha$. 在 $0<\alpha<1$ 时，倘若 $E_{\theta_1}\phi(X)=\alpha$，则 $\phi_1(x)\equiv\alpha$ 也是水平为 α 的 MPT. 由定理 3.2 知，则一定存在常数 $k\geqslant0$，使得 $\phi_1(x)$ 满足 (3.11) 式 (a.e. $[\mu]$). 而 $0<\alpha<1$，故在集合 $\{x: p(x;\theta_1)>0$ 或 $p(x;\theta_0)>0\}$ 上

$$p(x;\theta_1)=k\cdot p(x;\theta_0)\text{a.e.}[\mu]$$

由于 p 是概率密度，因此 $k=1$，即 $p(x;\theta_1)=p(x;\theta_0)$ a.e. $[\mu]$. 证毕.

§ 3.3　一致最优势检验

§ 3.3.1　一致最优势检验

关于简单原假设对简单备择假设的检验问题，N-P 基本引理给出了令人满意的回答. 但是，在实际问题中，往往出现的是复合假设的情况，这时最优的标准是什么？

定义 3.7　设 $(\mathscr{X},\mathscr{B},\mathscr{P}=\{P_\theta:\theta\in\Theta\})$ 为一参数统计结构，考虑检验问题 (Θ_0,Θ_1). 设 $\phi(x)$ 是水平为 α 的检验. 如果对任意一个水平为

α 的检验 $\phi_1(x)$，都有

$$E_\theta[\phi(X)] \geqslant E_\theta[\phi_1(X)], \quad \forall\, \theta \in \Theta_1$$

则称检验 $\phi(x)$ 是水平为 α 的一致最优势检验，记为 UMPT(Uniformly Most Powerful Test).

在某些特殊的情况下，UMPT 可以直接从 N-P 基本引理推出，但要利用下述的结论.

引理 3.4 设 $\phi(x)$ 是 $(\alpha,\Theta_0,\Theta_1)$ 检验，Θ_{01} 是 Θ_0 的子集. 如果 $\phi(x)$ 是 $(\alpha,\Theta_{01},\Theta_1)$ 的 UMPT，则 $\phi(x)$ 是 $(\alpha,\Theta_0,\Theta_1)$ 的 UMPT.

证明：必要性可从 UMPT 的定义看出. 现证其充分性.

如果 $\phi_1(x)$ 是任意一个 $(\alpha,\Theta_0,\Theta_1)$ 检验，则有 $E_\theta\phi_1(X) \leqslant \alpha, \theta \in \Theta_0$. 所以 $\phi_1(x)$ 也是 $(\alpha,\Theta_{01},\Theta_1)$ 检验. 由于 $\phi(x)$ 是 $(\alpha,\Theta_{01},\Theta_1)$ 的 UMPT，则应有

$$E_\theta[\phi(X)] \geqslant E_\theta[\phi_1(X)], \quad \forall\, \theta \in \Theta_1$$

这表明 $\phi(x)$ 是 $(\alpha,\Theta_0,\Theta_1)$ 的 UMPT. 证毕.

引理 3.5 设 $\phi(x)$ 是 $(\alpha,\Theta_0,\Theta_1)$ 检验. 则 $\phi(x)$ 是 $(\alpha,\Theta_0,\Theta_1)$ 的 UMPT 的充要条件是，对每一个 $\theta_1 \in \Theta_1$，$\phi(x)$ 是 $(\alpha,\Theta_0,\{\theta_1\})$ 的 MPT.

必要性和充分性可从 UMPT 的定义看出. 证明略.

定理 3.6 设 $\phi(x)$ 是 $(\alpha,\Theta_0,\Theta_1)$ 检验. 假如对某个 $\theta_0 \in \Theta_0$ 和对每一个 $\theta_1 \in \Theta_1$，$\phi(x)$ 都是 $(\alpha,\{\theta_0\},\{\theta_1\})$ 的 MPT，则 $\phi(x)$ 也是 $(\alpha,\Theta_0,\Theta_1)$ 的 UMPT.

这个定理是引理 3.4 和 3.5 的推论，证明略.

在例 3.2 中，由 N-P 基本引理已求得简单原假设 $H_0: \mu=0$ 对简单备择假设 $H_1: \mu=\mu_1 (\mu_1>0)$ 的水平为 α 的 MPT$\phi(x)$，其拒绝域为 $W=\{x: \bar{x} \geqslant \frac{U_{1-\alpha}}{\sqrt{n}}\}$. 这个检验与 μ_1 的具体数值无关，只要求 $\mu_1>0$. 由于原假设 $H_0: \mu=0$ 成立时，$E_0\phi(X)=\alpha$，并且由于 $\phi(x)$ 的势函数 $g(\mu)=E_\mu\phi(X)$ 是 μ 的严增函数，则在 $\mu \leqslant 0$ 时，$E_\mu\phi(X) \leqslant E_0\phi(X)=\alpha$. 所以 $\phi(x)$ 是原假设 $H_0: \mu \leqslant 0$ 对备择假设 $H_1: \mu>0$ 的水平为 α 的检验. 根据定理 3.6，这个检验 $\phi(x)$ 是 $H_0: \mu \leqslant 0$ 对 $H_1: \mu>0$ 的 UMPT.

这个例子告诉我们，如果简单原假设对简单备择假设的检验问题的 MPT 不依赖于备择假设的具体数值，则可适当扩大备择假设，而当势函数是单调函数时，也可适当扩大原假设，这样就可由 MPT 获得 UMPT. 这就扩大了 N-P 基本引理的应用范围. 例 3.4 也有类似的结果，它是例 3.1 考虑的检验问题的 UMPT. 例 3.3 的检验函数 $\phi(x)$（见(3.14)式）似依赖于 θ_1，事实上这个检验函数可等价地表示为

$$\phi(x)=\begin{cases}1, & x_{(n)}\geqslant \sqrt[n]{1-\alpha}\\ 0, & 0<x_{(n)}<\sqrt[n]{1-\alpha}\end{cases}$$

它与 θ_1 的具体数值无关，只要求 $\theta_1>1$ 就行了. 所以，与例 3.2，3.4 一样，例 3.3 也有类似的结果. 它是原假设 $H_0: \theta\leqslant 1$ 对备择假设 $H_1: \theta>1$ 的 UMPT.

一般来说，对于复合假设检验问题，MPT 的 $\phi(x)$ 依赖于备择假设中的 θ 值，则 UMPT 不一定存在. 那么在什么情况下存在呢？在什么情况下不存在呢？下面将引进单调似然比以及单参数指数型分布族的概念，然后就下列几种检验问题进行讨论：

(Ⅰ) 原假设 $H_0: \theta\leqslant\theta_0$ 对备择假设 $H_1: \theta>\theta_0$；

(Ⅱ) 原假设 $H_0: \theta\geqslant\theta_0$ 对备择假设 $H_1: \theta<\theta_0$；

(Ⅲ) 原假设 $H_0: \theta=\theta_0$ 对备择假设 $H_1: \theta\neq\theta_0$；

(Ⅳ) 原假设 $H_0: \theta_1\leqslant\theta\leqslant\theta_2$ 对备择假设 $H_1: \theta<\theta_1$ 或 $\theta>\theta_2$；

(Ⅴ) 原假设 $H_0: \theta\leqslant\theta_1$ 或 $\theta\geqslant\theta_2$ 对备择假设 $H_1: \theta_1<\theta<\theta_2$.

检验问题(Ⅰ)和(Ⅱ)称为单边假设检验问题，而检验问题(Ⅲ)、(Ⅳ)和(Ⅴ)称为双边假设检验问题. 首先寻求单边假设检验问题的 UMPT. 这需要引进单调似然比的概念.

§3.3.2　单调似然比

定义 3.8　设 $\{p(x;\theta): \theta\in\Theta\}$ 是含有实参数 θ 的概率密度族，其中 Θ 是实直线上的一个区间. 如果存在实值统计量 $T(X)$，使得对任意的 $\theta_1<\theta_2$，都有

(1) 概率分布 P_{θ_1} 与 P_{θ_2} 是不同的；

(2)似然比 $\lambda(x)=\dfrac{p(x;\theta_2)}{p(x;\theta_1)}$是 $T(x)$的非降函数(或非增函数),则称概率密度族$\{p(x;\theta):\theta\in\Theta\}$关于 $T(x)$具有非降(或非增)单调似然比.单调似然比简写为 MLR(Monotone Likelihood Ratio).

单参数指数型分布族

$$p(x;\theta)=c(\theta)\cdot\exp\{Q(\theta),T(x)\}\cdot h(x) \qquad (3.15)$$

其中 $c(\theta)>0$. 假设 $Q(\theta)$是 θ 的严增函数(或严减函数).则在 $\theta_1<\theta_2$ 时,其似然比

$$\lambda(x)=\frac{p(x;\theta_2)}{p(x;\theta_1)}=\frac{c(\theta_2)}{c(\theta_1)}\cdot\exp\{[Q(\theta_2)-Q(\theta_1)]\cdot T(x)\}$$

是 $T(x)$的严增函数(或严减函数).故单参数指数型分布族(3.15)关于 $T(x)$具有非降(或非增)MLR.这里的 $T(X)$是充分统计量.二项分布族,负二项分布族,泊松分布族,正态分布族(均值已知,方差未知或均值未知,方差已知的情况)和指数分布族,以及来自这些分布族的样本分布族都是单参数指数型分布族,它们关于其充分统计量都具有 MLR.虽然均匀分布族以及来自它的样本分布族不是单参数指数型分布族,但如果定义$\dfrac{0}{0}=\infty$,则均匀分布族以及来自它的样本分布族关于其充分统计量也都具有 MLR.

下面的定理揭示了 MLR 分布族的一个基本性质.

定理 3.7 设概率密度族$\{p(x;\theta):\theta\in\Theta\subseteq\mathbf{R}\}$关于 $T(x)$具有非降 MLR.若 $\psi(t)$是 t 的一个非降函数,则 $E_\theta\psi(T(X))$是 θ 的一个非降函数.

证明:设 $\theta_1<\theta_2$,令

$$A=\{x:p(x;\theta_1)<p(x;\theta_2)\},\quad B=\{x:p(x;\theta_1)>p(x;\theta_2)\}$$

对任意的 $x_1\in A$ 和 $x_2\in B$,有

$$\lambda(x_1)=\frac{p(x_1;\theta_2)}{p(x_1;\theta_1)}>1,\quad \lambda(x_2)=\frac{p(x_2;\theta_2)}{p(x_2;\theta_1)}<1$$

因为似然比 $\lambda(x)$是 $T(x)$的非降函数,则由 $\lambda(x_1)>\lambda(x_2)$可以推出 $T(x_1)\geqslant T(x_2)$.由于 $\psi(t)$是 t 的一个非降函数,所以 $\psi(T(x_1))\geqslant\psi(T(x_2))$.令

$$a=\inf\{\psi(T(x)):x\in A\},\quad b=\sup\{\psi(T(x)):x\in B\}$$

则有 $a\geqslant b$. 从而

$$\begin{aligned}
&\mathrm{E}_{\theta_2}\psi(T(X))-\mathrm{E}_{\theta_1}\psi(T(X))\\
&=\int\psi(T(x))\cdot[p(x;\theta_2)-p(x;\theta_1)]\mathrm{d}\mu(x)\\
&\geqslant a\cdot\int_A[p(x;\theta_2)-p(x;\theta_1)]\mathrm{d}\mu(x)+\\
&\quad b\cdot\int_B[p(x;\theta_2)-p(x;\theta_1)]\mathrm{d}\mu(x)
\end{aligned}\tag{3.16}$$

由于

$$\int_{A\cup B}[p(x;\theta_2)-p(x;\theta_1)]\mathrm{d}\mu(x)=0$$

所以

$$\int_B[p(x;\theta_2)-p(x;\theta_1)]\mathrm{d}\mu(x)=-\int_A[p(x;\theta_2)-p(x;\theta_1)]\mathrm{d}\mu(x)$$

从而

$$\begin{aligned}
&\mathrm{E}_{\theta_2}\psi(T(X))-\mathrm{E}_{\theta_1}\psi(T(X))\\
&\geqslant(a-b)\cdot\int_A[p(x;\theta_2)-p(x;\theta_1)]\mathrm{d}\mu(x)\geqslant 0
\end{aligned}$$

证毕.

推论 1°　若 $\psi(t)$是 t 的一个非增函数，则 $\mathrm{E}_\theta\psi(T(X))$是 θ 的一个非增函数;

推论 2°　如果概率密度族$\{p(x;\theta):\theta\in\Theta\subseteq\mathbf{R}\}$关于 $T(x)$具有非增 MLR，那么它关于$(-T(x))$具有非降 MLR. 由于在 $\psi(t)$是 t 的一个非降函数时，$\psi(t)$是$(-t)$的一个非增函数，所以 $\mathrm{E}_\theta\psi(T(X))$是 θ 的一个非增函数;而在 $\psi(t)$是 t 的一个非增函数时，$\mathrm{E}_\theta\psi(T(X))$是 θ 的一个非降函数.

在概率密度族$\{p(x;\theta):\theta\in\Theta\subseteq\mathbf{R}\}$关于 $T(x)$具有非降 MLR 时，取 $\psi(t)=t$，则在 $\theta_1<\theta_2$ 时

$$\mathrm{E}_{\theta_1}T(X)\leqslant\mathrm{E}_{\theta_2}T(X)\tag{3.17}$$

所以，在概率密度族关于 $T(x)$具有非降 MLR 时，$T(X)$的均值

$E_\theta T(X)$是θ的一个非降函数. 对任意给定的t_0,分别在$t>t_0$或$t\leqslant t_0$时,取$\psi(t)=0$或1. 则在$\theta_1<\theta_2$时,有

$$P_{\theta_1}(T(X)\leqslant t_0)\geqslant P_{\theta_2}(T(X)\leqslant t_0) \tag{3.18}$$

所以,在概率密度族关于$T(x)$具有非降MLR时,$T(X)$的分布函数$P_\theta(T(X)\leqslant t)$是$\theta$的一个非增函数.

譬如,随机变量$X\sim b(n,p)$,其中$0<p<1$为未知参数. 其概率密度为

$$p(x;p)=(1-p)^n\cdot\exp\left\{\ln\left(\frac{p}{1-p}\right)\cdot x\right\}\cdot\binom{n}{x}$$

这是单参数指数型分布族. 由于$Q(p)=\ln\left(\frac{p}{1-p}\right)$是$p$的严增函数,所以它关于$x$具有非降MLR. 二项分布的均值$E_pX=np$和分布函数$P_p(X\leqslant x)=\sum\limits_{k\leqslant x}\binom{n}{x}\cdot p^x\cdot(1-p)^{n-x}$分别是$p$的非降和非增函数. 事实上,它们分别是$p$的严增和严降函数.

又设$X=(X_1,\cdots,X_n)$是来自Poisson分布总体$\mathscr{P}=\{P(\lambda),\lambda>0\}$的一个样本. 样本联合概率密度为

$$p(x;\lambda)=e^{-n\lambda}\cdot\exp\left\{\ln\lambda\cdot\sum_{i=1}^n x_i\right\}\cdot\left(\prod_{i=1}^n(x_i!)\right)^{-1}$$

这是单参数指数型分布族. 由于$Q(\lambda)=\ln\lambda$是λ的严增函数,所以它关于$T(x)=\sum_{i=1}^n x_i$具有非降MLR. $T(X)\sim P(n\lambda)$. $T(X)$的均值$E_\lambda T(X)=n\lambda$和分布函数$P_\lambda(T(X)\leqslant t)=\sum\limits_{k\leqslant t}\frac{(n\lambda)^k}{k!}\cdot e^{-n\lambda}$分别是$\lambda$的非降和非增函数. 事实上,它们分别是$\lambda$的严增和严降函数.

在概率密度族关于$T(x)$具有非降MLR时,如果考虑单边假设检验问题(Ⅰ),则(3.17)和(3.18)两式皆表示,应在$T(x)$的值较大时拒绝原假设. 这个检验方法是符合人们直观感觉的,也是可信的. 下面我们将证明,它是UMPT.

§3.3.3 单边假设检验

定理3.8 设单参数概率密度族$\{p(x;\theta):\theta\in\Theta\subseteq\mathbf{R}\}$关于实值统

计量 $T(x)$ 具有非降 MLR，则关于单边假设检验问题(Ⅰ)，

(1)存在水平为 α 的 UMPT 的检验函数

$$\phi(T(x)) = \begin{cases} 1, & T(x) > c \\ r, & T(x) = c \\ 0, & T(x) < c \end{cases} \tag{3.19}$$

其中常数 $r(0 \leqslant r \leqslant 1)$ 和 c 由下式确定

$$\mathrm{E}_{\theta_0}\phi(T(X)) = \alpha \tag{3.20}$$

(2)这个检验的势函数

$$g(\theta) = \mathrm{E}_{\theta}\phi(T(X))$$

是非降的，且在集合 $\{\theta : 0 < g(\theta) < 1\}$ 上是严格增加的.

(3)在一切使得 $\mathrm{E}_{\theta_0}\phi(X) = \alpha$ 的检验函数中，由(3.19)和(3.20)两式所确定的检验函数 $\phi(T(x))$，使得对任意的 $\theta < \theta_0$，$\mathrm{E}_{\theta}\phi(X)$ 都达到最小.

证明：先证明(1). 考虑简单原假设 $H_0 : \theta = \theta_0$ 对简单备择假设 $H_1 : \theta = \theta_1 (\theta_1 > \theta_0)$ 的检验问题. 由 N-P 基本引理知，应在似然比统计量 $\lambda(x) = \dfrac{p(x;\theta_1)}{p(x;\theta_0)}$ 比较大的时候，拒绝原假设. 由于 $\lambda(x)$ 是 $T(x)$ 的非降函数，所以在 $T(x)$ 比较大的时候，拒绝原假设. 为此，我们采用形如(3.19)式的检验函数 $\phi(T(x))$. 用类似于定理 3.2 的证明方法可以知道，满足(3.19)和(3.20)两式的检验 $\phi(T(x))$ 是存在的. 具体地说，在 $T(x)$ 为连续型随机变量的时候，$\phi(T(x))$ 是非随机化检验，其拒绝域为 $W = \{T(x) \geqslant c\}$，c 由 $P_{\theta_0}\{T(X) \geqslant c\} = \alpha$ 确定. 在 $T(X)$ 为离散型随机变量的时候，c 由 $P_{\theta_0}\{T(X) > c\} \leqslant \alpha \leqslant P_{\theta_0}\{T(X) \geqslant c\}$ 确定，而 $r = \dfrac{\alpha - P_{\theta_0}\{T(X) > c\}}{P_{\theta_0}\{T(X) = c\}}$.

在 $T(x) = c$ 时，假设 $\lambda(x) = k$. 由于 $\lambda(x)$ 是 $T(x)$ 的非降函数，所以

$$\{x : \lambda(x) < k\} \subseteq \{x : T(x) < c\}$$
$$\{x : \lambda(x) > k\} \subseteq \{x : T(x) > c\}$$

因而满足(3.19)式的检验 $\phi(T(x))$ 必满足(3.11)式. 检验 $\phi(T(x))$ 满

足(3.20)式,也就是满足(3.10)式.从而由N P基本引理知,满足(3.19)和(3.20)两式的检验$\phi(T(x))$是该简单假设检验问题,$H_0:\theta=\theta_0$对$H_1:\theta=\theta_1(\theta_1>\theta_0)$的水平为$\alpha$的MPT.由于检验$\phi(T(x))$与$\theta_1$无关,只要求$\theta_1>\theta_0$,所以由引理3.5知,$\phi(T(x))$也是$H_0:\theta=\theta_0$对$H_1:\theta>\theta_0$的检验问题的水平为$\alpha$的UMPT.

由于$\phi(T(x))$关于$T(x)$非降,则据定理3.7可知,势函数$g(\theta)=E_\theta\phi(T(X))$是$\theta$的非降函数.故在$\theta\leqslant\theta_0$时,$E_\theta\phi(T(X))\leqslant E_{\theta_0}\phi(T(X))=\alpha$.所以对于原假设$H_0:\theta\leqslant\theta_0$,$\phi(T(x))$是水平为$\alpha$的检验.则据引理3.4可知,由(3.19)和(3.20)两式确定的检验函数$\phi(T(x))$是单边假设检验问题$H_0:\theta\leqslant\theta_0$对$H_1:\theta>\theta_0$的水平为$\alpha$的UMPT.结论(1)成立.

(2)的证明.我们已经证明了,势函数$g(\theta)=E_\theta\phi(T(X))$是$\theta$的非降函数.接下来证明,它在集合$\{\theta:0<g(\theta)<1\}$上是严格增加的.由N-P基本引理知,$\phi(T(x))$也是$H_0:\theta=\theta_1$对$H_1:\theta=\theta_2(\theta_2>\theta_1)$的水平为$\alpha_1=E_{\theta_1}\phi(T(X))$的MPT.则在$0<\alpha_1<1$时,由定理3.3知,$E_{\theta_2}\phi(T(X))>\alpha_1$.所以在集合$\{\theta:0<g(\theta)<1\}$上,势函数$g(\theta)=E_\theta\phi(T(X))$是严格增加的.结论(2)成立.

(3)的证明.考虑简单原假设$H_0:\theta=\theta_0$对简单备择假设$H_1:\theta=\theta_2(\theta_2<\theta_0)$的水平为$\alpha^*=1-\alpha$的检验问题.由N P基本引理知,我们在似然比统计量$\lambda(x)=\dfrac{p(x;\theta_2)}{p(x;\theta_0)}$比较大的时候,拒绝原假设.由于$\theta_2<\theta_0$,所以$\lambda(x)$是$T(x)$的非增函数.因此我们在$T(x)$比较小的时候,拒绝原假设.用类似于(1)的证明方法可以知道,$\phi^*(T(x))=1-\phi(T(x))$是$H_0:\theta=\theta_0$对$H_1:\theta=\theta_2(\theta_2<\theta_0)$的水平为$\alpha^*=1-\alpha$的MPT,其中$\phi(T(x))$由(3.19)和(3.20)两式给出.因此$\phi^*(T(x))$是在$E_{\theta_0}\phi(X)\leqslant\alpha^*$的条件下,使$E_{\theta_2}\phi(X)$达到最大.从而$\phi(T(x))$是在$E_{\theta_0}\phi(X)\geqslant\alpha$的条件下,使$E_{\theta_2}\phi(X)$达到最小.由于$\theta_2$是小于$\theta_0$的任意一个数,所以对任意的$\theta<\theta_0$,$\phi(T(x))$是在$E_{\theta_0}\phi(X)\geqslant\alpha$,当然也是在$E_{\theta_0}\phi(X)=\alpha$的条件下,使$E_\theta\phi(X)$达到最小的检验法.结论(3)成立.证毕.

关于这个定理有以下一些推论.

推论 1°　在定理 3.8 的条件下，考虑

(1)检验问题 $H_0: \theta=\theta_0$ 对 $H_1: \theta>\theta_0$，则定理 3.8 的结论仍全部成立；

(2)单边假设检验问题(Ⅰ)，则类似于定理 3.8 的结论全部成立，只需要将(3.19)式中的不等号改变方向；

(3)检验问题 $H_0: \theta=\theta_0$ 对 $H_1: \theta<\theta_0$，则推论 1°(2)的结论仍全部成立.

推论 2°　设单参数概率密度族 $\{p(x;\theta): \theta\in\Theta\subseteq\mathbf{R}\}$ 关于实值统计量 $T(x)$ 具有非增 MLR，从而它关于 $-T(x)$ 具有非降 MLR，定理 3.8 及其推论 1° 的结论仍全部成立，只需要将 $T(x)$ 换为 $-T(x)$.

推论 3°　单参数指数型分布族(3.14)，在 $Q(\theta)$ 是 θ 的严格单调函数时，它关于充分统计量 $T(x)$ 具有 MLR. 由定理 3.8 及其推论 1° 和 2° 可以知道，对于单参数指数型分布族(3.15)的未知参数 θ 的单边假设检验问题，都可以找到 UMPT.

推论 4°　设单参数概率密度族 $\{p(x;\theta): \theta\in\Theta\subseteq\mathbf{R}\}$ 关于实值统计量 $T(x)$ 具有非降 MLR，(3.18)式告诉我们 $T(X)$ 的分布函数 $P_\theta(T(X)\leqslant t)=1-P_\theta(T(X)>t)$ 是 θ 的一个非增函数. 现由定理 3.8 (2)可进一步推得，在集合 $\{\theta: 0<P_\theta(T(x)\leqslant t)<1\}$ 上，$T(X)$ 的分布函数是 θ 的一个严降函数. 如果单参数概率密度族关于实值统计量 $T(x)$ 具有非增 MLR，与上述相类似的结论仍然成立，只需要将关于 $T(X)$ 的分布函数的结论中的非增换为非降，严降换为严增. 这个结论将被用于第四章的置信区间(限)的问题.

例 3.5　若有 N 件产品，其中含有 m 件不合格品. 今从中不返回地随机抽取 n 件进行检验. 设 n 件中含有 x 件不合格品. x 的可能取值为 $0,1,2,\cdots\min(m,n)$. X 服从超几何分布. 当 m 未知时，考察超几何分布概率密度族 $\{p(x;m): m=0,1,2,\cdots,N\}$，其中

$$p(x;m)=\frac{\binom{m}{x}\cdot\binom{N-m}{n-x}}{\binom{N}{n}}$$

显然,此概率密度族不是单参数指数型分布族.因为

$$\frac{p(x;m+1)}{p(x;m)}=\frac{\binom{m+1}{x}\cdot\binom{N-m-1}{n-x}}{\binom{m}{x}\cdot\binom{N-m}{n-x}}=\frac{(m+1)(N-m-n+x)}{(n-m)(m+1-x)}$$

是 x 的严增函数,所以此概率密度族关于 $T(x)=x$ 具有 MLR.

考察如下的单边假设检验问题:

$$H_0: 0\leqslant m\leqslant m_0 \text{ 对 } H_1: m_0<m\leqslant N$$

由定理 3.8 知,此检验问题存在水平为 α 的 UMPT,其检验函数为

$$\phi(x)=\begin{cases}1, & x>c\\ \dfrac{(\alpha-\alpha_1)\cdot\binom{N}{n}}{\binom{m_0}{c}\cdot\binom{N-m_0}{n-c}}, & x=c\\ 0, & x<c\end{cases}$$

其中 c 满足条件

$$\sum_{x\geqslant c}\frac{\binom{m_0}{x}\cdot\binom{N-m_0}{n-x}}{\binom{N}{n}}>\alpha\geqslant\sum_{x\geqslant c+1}\frac{\binom{m_0}{x}\cdot\binom{N-m_0}{n-x}}{\binom{N}{n}}=\alpha$$

设有 $N=1000$ 件产品,其中含有 m 件不合格品.今从中不返回地随机抽取 $n=100$ 件进行检验.在 $m_0=10$ 时,考察上述的单边假设检验问题.取 $\alpha=0.05$.查阶乘对数表,得

$$p(0;10)=\frac{\binom{10}{0}\cdot\binom{990}{100}}{\binom{1000}{100}}=\frac{990!900!}{890!\cdot1000!}=0.3470$$

从而

$$p(1;10)=\frac{\binom{10}{1}\cdot\binom{990}{99}}{\binom{1000}{100}}=\frac{10\cdot 100}{891}\cdot p(0;10)=0.3894$$

$$p(2;10)=\frac{\binom{10}{2}\cdot\binom{990}{98}}{\binom{1000}{100}}=\frac{9\cdot 99}{2\cdot 892}\cdot p(1;10)=0.1945$$

$$p(3;10)=\frac{\binom{10}{3}\cdot\binom{990}{97}}{\binom{1000}{100}}=\frac{8\cdot 98}{3\cdot 893}\cdot p(2;10)=0.0569$$

则有

$$\sum_{x\geqslant 3}\frac{\binom{10}{x}\cdot\binom{990}{150-x}}{\binom{1000}{150}}=1-\sum_{x\leqslant 2}\frac{\binom{10}{x}\cdot\binom{990}{150-x}}{\binom{1000}{150}}$$

$$=0.0691>0.05$$

$$\sum_{x\geqslant 4}\frac{\binom{10}{x}\cdot\binom{990}{150-x}}{\binom{1000}{150}}=1-\sum_{x\leqslant 3}\frac{\binom{10}{x}\cdot\binom{990}{150-x}}{\binom{1000}{150}}$$

$$=0.0122<0.05$$

所以此检验问题的水平为 $\alpha=0.05$ 的 UMPT 为

$$\phi(x)=\begin{cases}1, & x\geqslant 4\\ 0.6643, & x=3\\ 0, & x\leqslant 2\end{cases}$$

例 3.6 设 $X=(X_1,\cdots,X_n)$是来自正态分布族$\{N(0,\sigma^2):\sigma^2>0\}$的样本. 求 $H_0:\sigma^2\leqslant\sigma_0^2$ 对 $H_1:\sigma^2>\sigma_0^2$ 的单边假设检验问题的水平为 α 的 UMPT. 样本的联合密度为

$$p(x;\sigma^2) = (\sqrt{2\pi}\sigma)^{-n} \cdot \exp\left\{(-2\sigma^2)^{-1} \cdot \sum_{i=1}^{n} x_i^2\right\}$$

其中 $Q(\sigma^2)=(-2\sigma^2)^{-1}$是 σ^2 的严增函数. 由定理 3.8 及其推论 3^0 知，其 UMPT 存在，且检验函数仅依赖于充分统计量 $T(X)=\sum_{i=1}^{n} X_i^2$，拒绝域为 $W=\{x: T(x)\geqslant c\}$，其中 c 由 $E_{\sigma_0^2}\phi(T(X))=\alpha$ 确定. 在 $\sigma^2=\sigma_0^2$ 时，$T(X)\sim\sigma_0^2\cdot\chi^2(n)$. 故 $c=\sigma_0^2\cdot\chi^2_{1-\alpha}(n)$.

取 $\sigma_0^2=1, n=10$ 和 $\alpha=0.1$. 该 UMPT 检验的拒绝域为

$$W = \left\{x: \sum_{i=1}^{10} x_i^2 \geqslant 15.99\right\}$$

其势函数 $g(\sigma^2)$的图形如图 3.2 上的粗曲线所示. 若令 $T_1(x)=\sum_{i=1}^{10}(x_i-\bar{x})^2$，取另一检验的拒绝域为 $W_1=\{x: T_1(x)\geqslant\sigma_0^2\cdot\chi^2_{1-\alpha}(n-1)\}$. 可以证明，这个检验也是该检验问题的水平为 α 的一个检验. $\sigma_0^2=1, n=10$ 和 $\alpha=0.1$ 时，其势函数 $g_1(\sigma^2)$的图形如图 3.2 上的细曲线所示. 在 $\sigma^2>1$ 时，粗曲线一致地在细曲线的上面，这正是定理 3.8 (1)所说的. 在 $\sigma^2=1$ 时，这两个检验的势函数值都是 0.1，而在 $\sigma^2<1$ 时，粗曲线一致地在细曲线的下面，这正是定理 3.8(3)所说的.

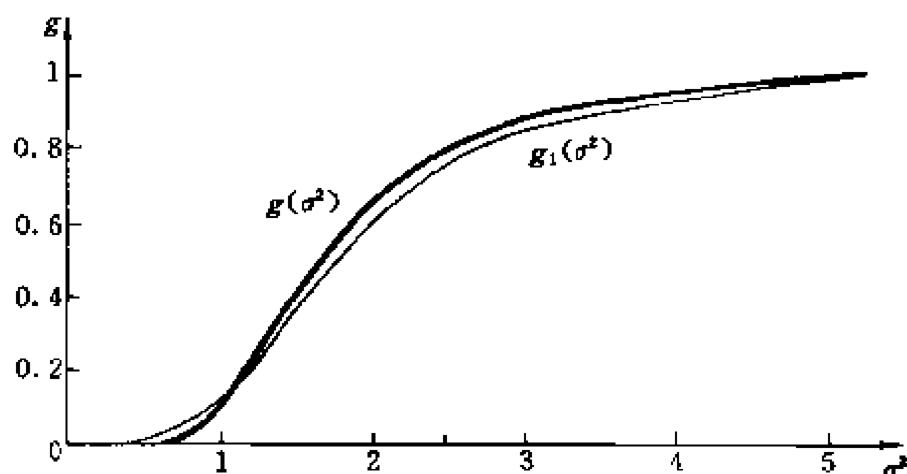


图 3.2　例 3.6 的两个检验的势函数图形的比较

§ 3.3.4　双边假设检验

对单参数指数型分布族，双边假设检验问题(Ⅴ)的 UMPT 是存

在的. 但一般来说,双边假设检验问题(Ⅲ)和(Ⅳ)的 UMPT 是不存在的. 看下面的例子.

例 3.7 设 $X=(X_1,\cdots,X_n)$为来自正态总体$\{N(\mu,1):-\infty<\mu<\infty\}$的一个样本. 考虑原假设 $H_0:\mu=0$ 对备择假设 $H_1:\mu\neq 0$ 的双边假设检验问题. 倘若 $\phi(x)$是它的水平为 α 的 UMPT. 由于 $\phi^*(x)\equiv\alpha$ 显然是该检验问题的水平为 α 的一个检验,则有

$$\mathrm{E}_\mu\phi(X)\geqslant \mathrm{E}_\mu\phi^*(X)=\alpha \qquad \mu\neq 0 \tag{3.21}$$

此外,据引理 3.5,$\phi(x)$也是 $H_0:\mu=0$ 对 $H_1:\mu=\mu_1(\mu_1>0)$的水平为 α 的 MPT. 则由例 3.2 知,$\phi(x)$是拒绝域为 $W=\left\{x:\overline{x}\geqslant\dfrac{U_{1-\alpha}}{\sqrt{n}}\right\}$的一个检验,其势函数 $g(\mu)=\mathrm{E}_\mu\phi(X)=\Phi(\sqrt{n}\cdot\mu-U_{1-\alpha})$是 μ 的严增函数,故在 $\mu<0$ 时,$\mathrm{E}_\mu\phi(X)<\mathrm{E}_0\phi(X)=\alpha$. 这与(3.21)式相矛盾. 所以,该双边假设检验问题不存在 UMPT. 同样,原假设 $H_0:\theta_1\leqslant\theta\leqslant\theta_2$ 对备择假设 $H_1:\theta<\theta_1$ 或 $\theta>\theta_2$ 的双边假设检验问题也不存在 UMPT. 这里的"不存在"是对所有的检验函数而言的. 如果我们对检验函数加上某种要求,例如无偏性,那么检验函数类就缩小了. 下一节,我们将说明,在缩小的检验函数类中存在 UMPT.

§3.3.5 N-P 基本引理的推广(一)

要解决双边假设检验问题(Ⅴ)的 UMPT 的存在性问题,先要对 N-P 基本引理进行推广.

定理 3.9 设 μ 为可测空间$(\mathscr{X},\mathscr{B})$上的测度. $p_0(x),p_1(x),\cdots,p_m(x)$为$(\mathscr{X},\mathscr{B})$上 $m+1$ 个实可测且 μ 可积函数. $\alpha_1,\cdots,\alpha_m$ 为给定的 m 个实常数. 如果存在检验函数 $\phi(x)$ 满足以下两个条件:

(1)若有 m 个非负常数 $k_1,\cdots,k_m$,使得

$$\phi(x)=\begin{cases}1, & p_0(x)>\sum\limits_{i=1}^{m}k_i\cdot p_i(x)\\[2ex] 0, & p_0(x)<\sum\limits_{i=1}^{m}k_i\cdot p_i(x)\end{cases}$$

(2) $\int\phi(x)\cdot p_i(x)\mathrm{d}\mu(x)=\alpha_i,i=1,\cdots,m.$

则对任一满足下式的检验函数 $\phi^*(x)$，

$$\int \phi^*(x) \cdot p_i(x) \mathrm{d}\mu(x) \leqslant \alpha_i, i = 1, \cdots, m \qquad (3.22)$$

都有

$$\int \phi(x) \cdot p_0(x) \mathrm{d}\mu(x) \geqslant \int \phi^*(x) \cdot p_0(x) \mathrm{d}\mu(x)$$

证明：定理的证明类似于定理 3.2 的(2)的证明. 显然有

$$\left[p_0(x) - \sum_{i=1}^{m} k_i \cdot p_i(x)\right] \cdot [\phi(x) - \phi^*(x)] \geqslant 0$$

于是

$$\int p_0(x) \cdot [\phi(x) - \phi^*(x)] \mathrm{d}\mu(x)$$

$$\geqslant \sum_{i=1}^{m} k_i \cdot \int p_i(x) \cdot [\phi(x) - \phi^*(x)] \mathrm{d}\mu(x)$$

从而

$$\int \phi(x) \cdot p_0(x) \mathrm{d}\mu(x) - \int \phi^*(x) \cdot p_0(x) \mathrm{d}\mu(x)$$

$$\geqslant \sum_{i=1}^{m} k_i \cdot \left[\alpha_i - \int \phi^*(x) \cdot p_i(x) \mathrm{d}\mu(x)\right] \geqslant 0 \qquad (3.23)$$

证毕.

在下一节，将进一步推广 N-P 基本引理. 这些定理中的结论实际上是一个充分条件. 类似于定理 3.2 的(1)和(3)的存在性和必要性也是可以证明的[1].

§ 3.3.6 单参数指数型分布族的双边假设检验问题(一)

定理 3.10 设样本 $X = (X_1, \cdots, X_n)$ 服从单参数指数族分布(3.15)，其中 θ 为实参数，$Q(\theta)$ 是 θ 的严增函数. 则关于双边假设检验问题(Ⅴ)，存在水平为 $\alpha(0<\alpha<1)$ 的 UMPT，其检验函数仅依赖于充分统计量 $T(x)$，形如

$$\phi(T(x))=\begin{cases}1, & c_1<T(x)<c_2\\ r_i, & T(x)=c_i, i=1,2\\ 0, & T(x)<c_1, \text{或} T(x)>c_2\end{cases} \tag{3.24}$$

其中常数 $r_i(0\leqslant r_i\leqslant 1)$和 c_i 由下式确定

$$E_{\theta_1}\phi(T(X))=E_{\theta_2}\phi(T(X))=\alpha \tag{3.25}$$

证明:从 $T(x)$是 θ 的充分统计量和定理 3.1 知,最优检验只要在由$T(x)$所构成的检验函数类中去寻找.对于单参数指数族分布,$T(x)$的导出分布族仍然是单参数指数族分布.考虑到 $Q(\theta)$是 θ 的严增函数.为简化起见,设 $T(x)$的密度函数为

$$p(t;\theta)=c(\theta)\cdot\exp\{\theta\cdot t\}\cdot h(t) \tag{3.26}$$

其中 $c(\theta)>0$.

先考虑原假设 $H_0:\theta=\theta_1$ 或 $\theta=\theta_2$ 对备择假设 $H_1:\theta=\theta_3$,$(\theta_1<\theta_3<\theta_2)$的检验问题.定理 3.9 启示我们,取检验函数形如

$$\phi(t)=\begin{cases}1, & p(t;\theta_3)>\sum_{i=1}^{2}k_i\cdot p(t;\theta_i)\\ 0, & p(t;\theta_3)<\sum_{i=1}^{2}k_i\cdot p(t;\theta_i)\end{cases} \tag{3.27}$$

并满足条件

$$\int\phi(x)\cdot p(t;\theta_1)\mathrm{d}x=\int\phi(x)\cdot p(t;\theta_2)\mathrm{d}x=\alpha \tag{3.28}$$

(3.27)式化简后得到

$$\phi(t)=\begin{cases}1, & a_1\cdot e^{b_1t}+a_2\cdot e^{b_2t}<1\\ 0, & a_1\cdot e^{b_1t}+a_2\cdot e^{b_2t}>1\end{cases}$$

其中 $a_i=\frac{k_i\cdot c(\theta_i)}{c(\theta_3)}$,$b_i=\theta_i-\theta_3$,$i=1,2$. 则 $b_1<0<b_2$.

倘若 a_1 和 a_2 都是负的,则检验总是拒绝的,其势函数恒等于 1,这与水平 $\alpha(0<\alpha<1)$的条件相矛盾.倘若 a_1 和 a_2 中,一个是负的,而另一个是正的,或 a_1 和 a_2 中有一个为 0,则 $a_1\cdot e^{b_1t}+a_2\cdot e^{b_2t}$是严格单调的,所得到的检验如(3.19)式所示(见定理 3.8,单边假设检验),其势函数是单调函数,它不可能满足(3.28)式.所以,a_1 和 a_2 必定都是正

的. 从而 k_1 和 k_2 也都是正的. 故定理 3.9 的一个条件:"诸 k_i 都是非负的"得到满足.

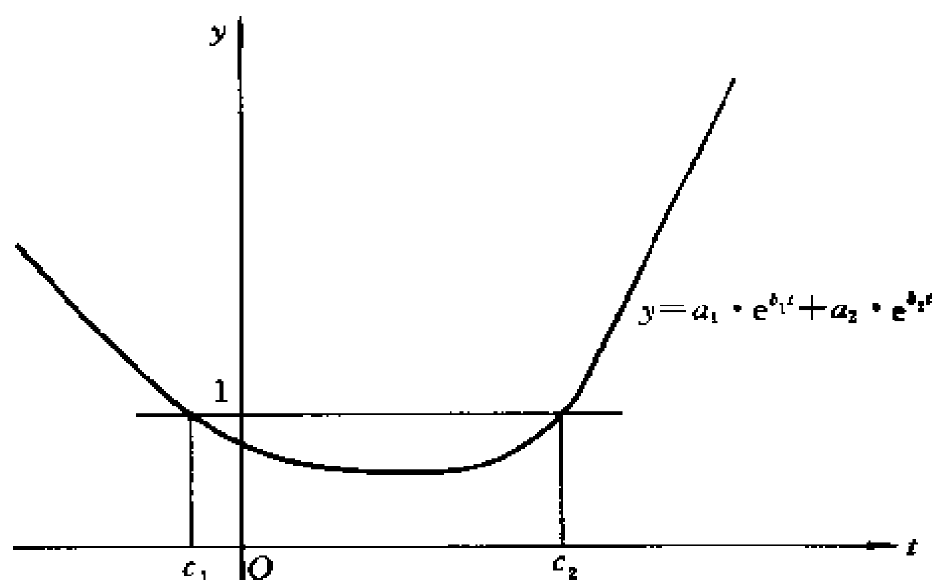


图 3.3 $y=a_1 \cdot e^{b_1 t}+a_2 \cdot e^{b_2 t}(b_1<0<b_2, a_1, a_2>0)$的图形

由于 $b_1<0<b_2$,并且 a_1 和 a_2 都是正的,所以在 $t\to\pm\infty$时,$y=a_1 \cdot e^{b_1 t}+a_2 \cdot e^{b_2 t}\to+\infty$,并且 $y''>0$. 故曲线 $y=a_1 \cdot e^{b_1 t}+a_2 \cdot e^{b_2 t}$是严格下凸的(见图 3.3). 则存在常数 c_1 和 c_2,使得(3.24)式成立. 因此,由(3.27)和(3.28)两式确定的检验函数就是由(3.24)和(3.25)两式确定的检验函数.

因为 k_1 和 k_2 都是正的,则据定理 3.9 知,由(3.24)和(3.25)两式确定的检验 $\phi(T(x))$(也就是 $\phi(t)$)是原假设 $H_0: \theta=\theta_1$ 或 $\theta=\theta_2$ 对备择假设 $H_1: \theta=\theta_3(\theta_1<\theta_3<\theta_2)$的检验问题的 MPT. 因为 $\phi(t)$与 θ_3 无关,只要求 $\theta_1<\theta_3<\theta_2$ 就行了,所以它也是原假设 $H_0: \theta=\theta_1$ 或 $\theta=\theta_2$ 对备择假设 $H_1: \theta_1<\theta<\theta_2$ 的检验问题的 UMPT.

为了应用引理 3.4,完成定理的证明,即证明 $\phi(t)$是原假设 $H_0: \theta\leqslant\theta_1$ 或 $\theta\geqslant\theta_2$ 对备择假设 $H_1: \theta_1<\theta<\theta_2$ 的检验问题的 UMPT,尚需证明,对于原假设 $H_0: \theta\leqslant\theta_1$ 或 $\theta\geqslant\theta_2$ 而言,$\phi(t)$是水平为 α 的一个检验. 我们将把这个证明留待下一节.

§3.4 一致最优势无偏检验

§3.4.1 无偏检验

在上节末尾，我们结合正态分布族$\{N(\mu,1):-\infty<\mu<\infty\}$指出，原假设$H_0:\mu=0$对备择假设$H_1:\mu\neq 0$的双边假设检验问题(Ⅱ)不存在UMPT.下面对这个例子作进一步的分析.任意选取一个检验函数$\phi(x)$，其对应的势函数$g(\mu)=E_\mu\phi(X)$是参数μ的函数.当然，我们希望在备择假设$H_1:\mu\neq 0$成立时，$E_\mu\phi(X)$，越大越好.因此，很自然地要求选取的检验函数$\phi(x)$的势函数$E_\mu\phi(X)$除了具有性质：$E_0\phi(X)\leqslant\alpha$外，还应具有性质：$E_\mu\phi(X)\geqslant\alpha,\mu\neq 0$.这样的$\phi(x)$就称为无偏检验函数.如果在所有的检验函数中无法找到UMPT，则可考虑在无偏检验函数类中去找UMPT.

定义3.9 设$\phi(x)$是原假设$H_0:\theta\in\Theta_0$对备择假设$H_1:\theta\in\Theta_1$的检验函数，若其势函数$g(\theta)=E_\theta\phi(X)$满足条件：

$$g(\theta)\leqslant\alpha,\quad\forall\,\theta\in\Theta_0$$
$$g(\theta)\geqslant\alpha,\quad\forall\,\theta\in\Theta_1$$

则称$\phi(x)$为水平为α的无偏检验.

显然，水平为α的UMPT一定是水平为α的无偏检验.

定义3.10 在检验问题(Θ_0,Θ_1)中，如果存在一个水平为α的无偏检验$\phi(x)$，对于任何水平为α的无偏检验$\phi_1(x)$，下列条件均满足

$$E_\theta[\phi(X)]\geqslant E_\theta[\phi_1(X)],\theta\in\Theta_1$$

则称检验$\phi(x)$是水平为α的一致最优势无偏检验，记为UMPUT (Uniformly Most Powerful Unbiased Test).

§3.4.2 相似检验

如果无偏检验$\phi(x)$的势函数$E_\theta\phi(X)$是θ的连续函数，则一定有

$$E_\theta\phi(X)=\alpha,\quad\theta\in\Gamma\tag{3.29}$$

其中Γ是Θ_0和Θ_1的公共边界.

定义 3.11 称满足(3.29)式的检验 $\phi(x)$ 为边界相似检验(Similar Test).

令 $\Phi^{(\alpha)}$,$\Phi_u^{(\alpha)}$ 和 $\Phi_s^{(\alpha)}$ 分别表示所有水平为 α 的检验,所有水平为 α 的无偏检验和所有的满足(3.29)式的边界相似检验构成的类. 显然,$\Phi_u^{(\alpha)}\subseteq\Phi^{(\alpha)}$,并且在势函数是连续函数时,$\Phi_u^{(\alpha)}\subseteq\Phi_s^{(\alpha)}$. 它们之间的关系如图 3.4 所示. 其中 Φ 表示所有的检验函数构成的类.

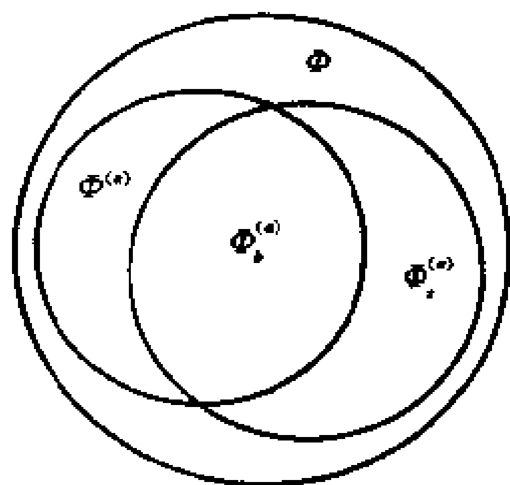


图 3.4 各种检验函数类之间的关系

一般来说,构造边界相似检验比构造无偏检验更为方便. 下面的引理在构造 UMPUT 方面是有意义的.

引理 3.11 考虑原假设 $H_0:\theta\in\Theta_0$ 对备择假设 $H_1:\theta\in\Theta_1$ 的检验问题. 设每一个检验函数的势函数皆为 θ 的连续函数. 如果 $\phi(x)$ 是水平为 α 的检验函数,则在 $\phi(x)$ 是所有满足(3.29)式的边界相似检验函数类中的一致最优势检验(UMPT)时,$\phi(x)$ 是水平为 α 的 UMPUT.

证明:显然,$\phi_1(x)\equiv\alpha$ 是边界相似的检验,故

$$\mathrm{E}_\theta[\phi(X)]\geqslant\mathrm{E}_\theta[\phi_1(X)]=\alpha,\quad\forall\,\theta\in\Theta_1$$

所以 $\phi(x)$ 是无偏检验. 由于满足(3.29)式的检验函数类 $\Phi_s^{(\alpha)}$ 包含了水平为 α 的无偏检验函数类 $\Phi_u^{(\alpha)}$,故 $\phi(x)$ 是水平为 α 的 UMPUT. 证毕.

由定理 3.8 的证明过程知,MLR 分布族的单边假设检验问题的 UMPT 就是一致最优势边界相似检验.

§3.4.3 N-P 基本引理的推广(二)

要解决双边假设检验(Ⅲ)和(Ⅳ)(见§3.3.3)的 UMPUT 问题,需要对 N-P 基本引理作进一步的推广.

定理 3.12 定理的条件和 $\phi(x)$ 同定理 3.9,只是常数 $k_1,\cdots,k_m$ 可以取负值. 则对任一满足下式

$$\int \phi^*(x) p_i(x) \mathrm{d}\mu(x) = \alpha_i, i = 1, \cdots, m$$

的检验函数 $\phi^*(x)$，都有

$$\int \phi(x) p_0(x) \mathrm{d}\mu(x) \geqslant \int \phi^*(x) p_0(x) \mathrm{d}\mu(x)$$

定理的证明完全类似于定理 3.9，只是(3.23)式中的最后一个不等号要改为等号. 证毕.

同定理 3.9 后的说明，这些定理的结论实际上是一个充分条件. 类似于定理 3.2 的(1)和(3)的存在性和必要性也是可以证明的[1].

现回到 §3.3.6 中定理 3.10 的证明. 我们已经证明了，存在满足(3.24)和(3.25)两式的检验函数 $\phi(t)$，它是原假设 $H_0: \theta=\theta_1$ 或 $\theta=\theta_2$ 对备择假设 $H_1: \theta_1<\theta<\theta_2$ 的检验问题的 UMPT. 为了应用引理 3.4，完成定理的证明，即证明 $\phi(t)$ 是 $H_0: \theta\leqslant\theta_1$ 或 $\theta\geqslant\theta_2$ 对 $H_1: \theta_1<\theta<\theta_2$ 的检验问题的 UMPT，尚需证明，对于原假设 $H_0: \theta\leqslant\theta_1$ 或 $\theta\geqslant\theta_2$ 而言，$\phi(t)$ 是水平为 α 的一个检验. 为此任取 $\theta'<\theta_1$（若取 $\theta'>\theta_2$，可同样处理），考虑原假设 $H_0: \theta=\theta_1$ 或 $\theta=\theta_2$ 对备择假设 $H_1: \theta=\theta'$ 的检验问题. 如果存在 k'_1 和 k'_2，使得

$$1-\phi(t)=\begin{cases}1, & p(t;\theta') > \sum_{i=1}^{2} k'_i \cdot p(t;\theta_i) \\ 0, & p(t;\theta') < \sum_{i=1}^{2} k'_i \cdot p(t;\theta_i)\end{cases} \tag{3.30}$$

则由 $\mathrm{E}_{\theta_1}[1-\phi(T)]=\mathrm{E}_{\theta_2}[1-\phi(T)]=1-\alpha$，以及定理 3.12 知

$$\mathrm{E}_{\theta'}[1-\phi(T)] \geqslant \mathrm{E}_{\theta'}\phi^*(T) = 1-\alpha$$

其中，$\phi^*(t)\equiv 1-\alpha$. 从而 $\mathrm{E}_{\theta'}\phi(T)\leqslant\alpha$. 这说明，对于原假设 $H_0: \theta\leqslant\theta_1$ 或 $\theta\geqslant\theta_2$ 而言，$\phi(t)$ 是水平为 α 的一个检验. 若找到使得(3.30)式成立的 k'_1 和 k'_2，则定理 3.10 就证毕了.

下面寻找使得(3.30)式成立的 k'_1 和 k'_2. 由(3.24)知

$$1-\phi(t)=\begin{cases}1, & t<c_1, \text{或 } t>c_2 \\ 0, & c_1<t<c_2\end{cases}$$

把它与(3.30)式进行比较. 由此可见，k'_1 和 k'_2 应满足条件：

$$p(t;\theta') > \sum_{i=1}^{2} k'_i \cdot p(t;\theta_i) \Leftrightarrow t < c_1 \text{ 或 } t > c_2 \tag{3.31}$$

$$p(t;\theta') < \sum_{i=1}^{2} k'_i \cdot p(t;\theta_i) \Leftrightarrow c_1 < t < c_2 \tag{3.32}$$

所以在 $t=c_1$ 和 $t=c_2$ 时，$p(t;\theta') = \sum_{i=1}^{2} k'_i \cdot p(t;\theta_i)$. 从而有下列方程组

$$\begin{cases} p(c_1;\theta') = k'_1 \cdot p(c_1;\theta_1) + k'_2 \cdot p(c_1;\theta_2) \\ p(c_2;\theta') = k'_1 \cdot p(c_2;\theta_1) + k'_2 \cdot p(c_2;\theta_2) \end{cases}$$

其中 c_1 和 c_2 已知，k'_1 和 k'_2 未知. 解此方程组，得到 k'_1 和 k'_2 的值. 它们分别是

$$\begin{cases} k'_1 = \dfrac{c(\theta')}{c(\theta_1)} \cdot \dfrac{\exp\{\theta' \cdot c_1 + \theta_2 \cdot c_2\} - \exp\{\theta' \cdot c_2 + \theta_2 \cdot c_1\}}{\exp\{\theta_1 \cdot c_1 + \theta_2 \cdot c_2\} - \exp\{\theta_1 \cdot c_2 + \theta_2 \cdot c_1\}} \\ k'_2 = \dfrac{c(\theta')}{c(\theta_2)} \cdot \dfrac{\exp\{\theta_1 \cdot c_1 + \theta' \cdot c_2\} - \exp\{\theta_1 \cdot c_2 + \theta' \cdot c_1\}}{\exp\{\theta_1 \cdot c_1 + \theta_2 \cdot c_2\} - \exp\{\theta_1 \cdot c_2 + \theta_2 \cdot c_1\}} \end{cases} \tag{3.33}$$

由于 $\theta' < \theta_1 < \theta_2$ 和 $c_1 < c_2$，所以

$$(\theta_1 \cdot c_1 + \theta_2 \cdot c_2) - (\theta_1 \cdot c_2 + \theta_2 \cdot c_1) = (\theta_1 - \theta_2) \cdot (c_1 - c_2) > 0$$

$$(\theta' \cdot c_1 + \theta_2 \cdot c_2) - (\theta' \cdot c_2 + \theta_2 \cdot c_1) = (\theta' - \theta_2) \cdot (c_1 - c_2) > 0$$

$$(\theta_1 \cdot c_1 + \theta' \cdot c_2) - (\theta_1 \cdot c_2 + \theta' \cdot c_1) = (\theta_1 - \theta') \cdot (c_1 - c_2) < 0$$

因此 $k'_1 > 0, k'_2 < 0$. 令

$$y = \frac{k'_1 \cdot p(t;\theta_1) + k'_2 \cdot p(t;\theta_2)}{p(t;\theta')} = a_1 \cdot e^{b_1 t} + a_2 \cdot e^{b_2 t}$$

其中 $a_1 = k'_1 \cdot \dfrac{c(\theta_1)}{c(\theta')} > 0, a_2 = k'_2 \cdot \dfrac{c(\theta_2)}{c(\theta')} < 0, b_1 = \theta_1 - \theta' > 0, b_2 = \theta_2 - \theta' > 0$，且 $b_1 < b_2$. 在 $t=c_1$ 或 c_2 时，$y = a_1 \cdot e^{b_1 t} + a_2 \cdot e^{b_2 t} = 1$.

曲线 $y = a_1 \cdot e^{b_1 t} + a_2 \cdot e^{b_2 t}$ 是单峰的. 自峰顶起向右（即当 $t \to +\infty$ 时），$y \to -\infty$；向左（即当 $t \to -\infty$ 时），$y \to 0$. $y = a_1 \cdot e^{b_1 t} + a_2 \cdot e^{b_2 t}$ $(0 < b_1 < b_2, a_1 > 0, a_2 < 0)$ 的图形由图 3.5 给出. 由此可见，

$$p(t;\theta') > \sum_{i=1}^{2} k'_i \cdot p(t;\theta_i) \Leftrightarrow 1 > a_1 \cdot e^{b_1 t} + a_2 \cdot e^{b_2 t} \Leftrightarrow t < c_1 \text{ 或 } t > c_2$$

$$p(t;\theta')<\sum_{i=1}^{2}k'_i\cdot p(t;\theta_i)\Leftrightarrow 1<a_1\cdot e^{b_1t}+a_2\cdot e^{b_2t}\Leftrightarrow c_1<t<c_2$$

所以，使得(3.31)和(3.32)两式，即使得(3.30)式成立的 k'_1 和 k'_2 找到了，如(3.33)式所示. 定理 3.10 证毕.

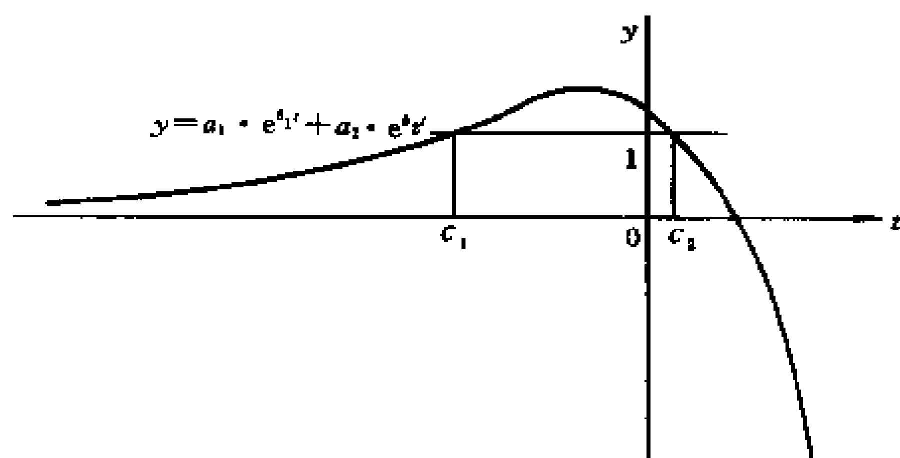


图 3.5 $y=a_1\cdot e^{b_1t}+a_2\cdot e^{b_2t}(0<b_1<b_2,a_1>0,a_2<0)$的图形

§3.4.4 单参数指数型分布族的双边假设检验问题(二)

定理 3.13 设样本 $X=(X_1,\cdots,X_n)$服从单参数指数族分布(3.15)，其中 θ 为实参数，$Q(\theta)$是 θ 的严增函数. 则关于原假设 $H_0:\theta_1\leqslant\theta\leqslant\theta_2$ 对备择假设 $H_1:\theta<\theta_1$ 或 $\theta>\theta_2$ 的双边假设检验问题(Ⅳ)，存在水平为 $\alpha(0<\alpha<1)$的 UMPUT，其检验函数为

$$\phi(T(x))=\begin{cases}1, & T(x)<c_1\text{ 或 }T(x)>c_2\\ r_i, & T(x)=c_i,i=1,2\\ 0, & c_1<T(x)<c_2\end{cases}\tag{3.34}$$

其中常数 $r_i(0\leqslant r_i\leqslant 1)$和 c_i 由下式确定

$$E_{\theta_1}\phi(T(X))=E_{\theta_2}\phi(T(X))=\alpha\tag{3.35}$$

证明：同定理 3.10，我们仅在由 $T(x)$所构造的检验函数类中去找最优检验. $T(x)$的分布密度函数是(3.26). 先考虑原假设 $H_0:\theta=\theta_1$ 或 $\theta=\theta_2$ 对备择假设 $H_1:\theta=\theta_3$，其中 $\theta_3<\theta_1<\theta_2$(若 $\theta_1<\theta_2<\theta_3$，可同样

处理). 受定理 3.9 和 3.12 的启示,取检验函数形如

$$\phi(t)=\begin{cases}1, & p(t;\theta_3)>\sum_{i=1}^{2}k_i\cdot p(t;\theta_i)\\ 0, & p(t;\theta_3)<\sum_{i=1}^{2}k_i\cdot p(t;\theta_i)\end{cases} \tag{3.36}$$

并满足条件(3.35)式. 化简后得

$$\phi(t)=\begin{cases}1, & a_1\cdot e^{b_1t}+a_2\cdot e^{b_2t}<1\\ 0, & a_1\cdot e^{b_1t}+a_2\cdot e^{b_2t}>1\end{cases}$$

其中 $a_i=\dfrac{k_i\cdot c(\theta_i)}{c(\theta_3)}$, $b_i=\theta_i-\theta_3$, $i=1,2$. 则 $0<b_1<b_2$.

倘若 a_1 和 a_2 都是负的,则检验总是拒绝的,其势函数恒等于 1,这与水平为 $\alpha(0<\alpha<1)$ 的条件相矛盾. 倘若 a_1 和 a_2 都是正的,或 a_1 和 a_2 中有一个为 0,则 $a_1\cdot e^{b_1t}+a_2\cdot e^{b_2t}$ 是严格单调的,所得到的检验如(3.18)式所示(见定理 3.8,单边假设检验),其势函数是单调函数,它不可能满足(3.35)式. 所以在 a_1 和 a_2 中,一个是负的,而另一个是正的. 倘若 $a_1<0, a_2>0$,那么在 $y=a_1\cdot e^{b_1t}+a_2\cdot e^{b_2t}>0$ 时,由于 $a_1\cdot e^{b_1t}>-a_2\cdot e^{b_2t}$,且 $0<b_1<b_2$,则

$$\begin{aligned}y'&=a_1\cdot b_1\cdot e^{b_1t}+a_2\cdot b_2\cdot e^{b_2t}\\ &>b_1\cdot(-a_2\cdot e^{b_2t})+a_2\cdot b_2\cdot e^{b_2t}=a_2\cdot e^{b_2t}\cdot(-b_1+b_2)>0\end{aligned}$$

这就是说,在 t 轴的上方,曲线 $y=a_1\cdot e^{b_1t}+a_2\cdot e^{b_2t}$ 是上升的. 所得到的检验也如(3.19)式所示,其势函数是单调函数,该检验不可能满足(3.35)式. 所以,a_1 必定是正的,而 a_2 必定是负的. 从而 $k_1>0, k_2<0$. 这就使得我们无法使用定理 3.9 去得到 UMPT,而只能使用定理 3.12 去得到 UMPUT.

在 $0<b_1<b_2, a_1>0, a_2<0$ 时,根据函数 $y=a_1\cdot e^{b_1t}+a_2\cdot e^{b_2t}$ 的图形可以知道,存在常数 c_1 和 c_2,使得(3.34)式成立. 因此,由(3.34)和(3.35)两式确定的检验函数就是由(3.36)和(3.35)两式确定的检验函数. 据定理 3.12,由(3.34)和(3.35)两式确定的检验函数 $\phi(T(x))$(也就是 $\phi(t)$)具有这样一个性质:对任一满足 $E_{\theta_1}\phi^*(T)=E_{\theta_2}\phi^*(T)=\alpha$ 的检验函数 $\phi^*(t)$,都有 $E_{\theta_3}\phi(T)\geqslant E_{\theta_3}\phi^*(T)$. 由于检验 $\phi(t)$ 与 θ_3 无关,只

要求 $\theta_3<\theta_1<\theta_2$，或 $\theta_1<\theta_2<\theta_3$ 就行了，所以由(3.34)和(3.35)两式确定的检验 $\phi(t)$ 是原假设 $H_0: \theta_1\leqslant\theta\leqslant\theta_2$ 对备择假设 $H_1: \theta<\theta_1$ 或 $\theta>\theta_2$ 的双边假设检验问题的一致最优势边界相似的检验. 为了应用引理 3.11 完成定理的证明，只需证明：对于原假设 $H_0: \theta_1\leqslant\theta\leqslant\theta_2$ 而言，$\phi(t)$ 是水平为 α 的一个检验. 运用与证明定理 3.10 相应结论完全类似的方法与步骤，可以得到其证明(证明过程从略). 定理证毕.

定理 3.14 设样本 $X=(X_1,\cdots,X_n)$ 服从单参数指数族分布(3.15)，其中 θ 为实值参数，$Q(\theta)$ 是 θ 的严格增加函数. 则关于原假设 $H_0: \theta=\theta_0$ 对备择假设 $H_1: \theta\neq\theta_0$ 的双边假设检验问题(Ⅱ)，存在水平为 $\alpha(0<\alpha<1)$ 的 UMPUT，其检验函数由(3.34)式给出，其中的常数 $r_i(0\leqslant r_i\leqslant 1)$ 和 c_i 由以下两式确定

$$E_{\theta_0}\phi(T(X))=\alpha \tag{3.37}$$

$$E_{\theta_0}[T(X)\cdot\phi(T(X))]=\alpha\cdot E_{\theta_0}T(X) \tag{3.38}$$

证明：同定理 3.14，我们仅在由 $T(x)$ 所构造的检验函数类中去找最优检验. $T(x)$ 的分布密度函数是(3.26).

首先证明，原假设 $H_0: \theta=\theta_0$ 对备择假设 $H_1: \theta\neq\theta_0$ 的任意一个无偏检验 $\phi(t)$ 必满足(3.37)和(3.38)两式. 显然，由检验函数 $\phi(t)$ 的无偏性可知，其势函数 $g(\theta)=E_\theta\phi(T)$ 在 $\theta=\theta_0$ 处达到最小值 α. 故 $g(\theta_0)=\alpha, g'(\theta_0)=0$. 由 $g(\theta_0)=\alpha$ 得(3.37)式. 由于 $g(\theta)=E_\theta\phi(T)=\int\phi(t)\cdot c(\theta)\cdot\exp\{\theta\cdot t\}\cdot h(t)d\mu(t)$，所以

$$\begin{aligned}g'(\theta)=&\int\phi(t)\cdot c'(\theta)\cdot\exp\{\theta\cdot t\}\cdot h(t)d\mu(t)+\\&\int\phi(t)\cdot c(\theta)\cdot t\cdot\exp\{\theta\cdot t\}\cdot h(t)d\mu(t)\\=&\frac{c'(\theta)}{c(\theta)}\cdot E_\theta[\phi(T)]+E_\theta[T\cdot\phi(T)]\end{aligned} \tag{3.39}$$

由于 $\frac{c'(\theta)}{c(\theta)}=-E_\theta[T]$(见习题 3.10)，则有

$$g'(\theta)=-E_\theta[T]\cdot E_\theta[\phi(T)]+E_\theta[T\cdot\phi(T)]$$

从而由 $g'(\theta_0)=0$，以及(3.37)式，得 $0=-E_\theta[T]\cdot\alpha+E_\theta[T\cdot\phi(T)]$，

则有(3.38)式.

接下来考虑原假设 $H_0: \theta=\theta_0$ 对备择假设 $H_1: \theta=\theta_1$,($\theta_1\neq\theta_0$)的检验问题,并要求检验函数 $\phi(t)$满足(3.37)和(3.38),即以下两式

$$\int\phi(t)\cdot p(t;\theta_0)\mathrm{d}\mu(t)=\alpha$$

$$\int\phi(t)\cdot t\cdot p(t;\theta_0)\mathrm{d}\mu(t)=\alpha\cdot \mathrm{E}_{\theta_0}[T]$$

定理 3.12 启示我们,取检验函数形如

$$\phi(t)=\begin{cases}1, & p(t;\theta_1)>k_1\cdot p(t;\theta_0)+k_2\cdot t\cdot p(t;\theta_0)\\ 0, & p(t;\theta_1)<k_1\cdot p(t;\theta_0)+k_2\cdot t\cdot p(t;\theta_0)\end{cases} \tag{3.40}$$

并满足条件(3.37)和(3.38)两式. 化简后得

$$\phi(t)=\begin{cases}1, & a_1+a_2\cdot t<\mathrm{e}^{bt}\\ 0, & a_1+a_2\cdot t>\mathrm{e}^{bt}\end{cases}$$

其中 $a_i=\dfrac{k_i\cdot c(\theta_0)}{c(\theta_i)}$,$i=1,2$,$b=\theta_1-\theta_0$.

上述不等式两边,右边是指数函数,左边是线性函数. 倘若指数曲线 $y=\mathrm{e}^{bt}$与直线 $y=a_1+a_2\cdot t$ 没有交点,则指数曲线在直线的上方. 故检验总是拒绝的,其势函数恒等于 1,这与水平 $\alpha(0<\alpha<1)$的条件相矛盾. 倘若指数曲线与直线相交一点,则所得到的检验如(3.19)式所示(见定理 3.8,单边假设检验),其势函数是严格单调函数,它不可能满足(3.38)式(为什么?见习题 3.11). 所以指数曲线与直线相交两点,即存在常数 c_1 和 c_2,使得(3.34)式成立. 因此,由(3.40),(3.37)和(3.38)三式确定的检验函数就是由(3.34),(3.37)和(3.38)三式确定的检验函数.

由(3.34),(3.37)和(3.38)三式确定的检验函数 $\phi(T(x))$(也就是 $\phi(t)$)与 θ_1 无关,只要求 $\theta_1\neq\theta_0$ 就行了. 据定理 3.12,该检验函数 $\phi(t)$ 具有这样一个性质:对任一满足(3.37)和(3.38)两式的检验函数 $\phi^*(t)$,都有 $\mathrm{E}_{\theta_1}\phi(T)\geqslant\mathrm{E}_{\theta_1}\phi^*(T)$. 若取 $\phi^*(t)=\alpha$,它满足(3.37)和(3.38)两式. 则由 $\phi(t)$具有的这样一个性质,有 $\mathrm{E}_{\theta_1}\phi(T)\geqslant\mathrm{E}_{\theta_1}\phi^*(T)=\alpha$,$\forall\ \theta_1\neq\theta_0$. 所以该检验函数 $\phi(t)$是原假设 $H_0: \theta=\theta_0$ 对备择假设 $H_1: \theta\neq\theta_0$ 的一个无偏检验,其势函数 $g(\theta)=\mathrm{E}_\theta\phi(T)$在 $\theta=\theta_0$ 处达到最小

值 α. 又取 $\phi^*(t)$ 为原假设 $H_0: \theta=\theta_0$ 对备择假设 $H_1: \theta\neq\theta_0$ 的任意一个无偏估计，它满足(3.37)和(3.38)两式，则由 $\phi(t)$ 具有的这样一个性质，有 $E_{\theta_1}\phi(T)\geqslant E_{\theta_1}\phi^*(T)$, $\forall\ \theta_1\neq\theta_0$. 所以该检验函数 $\phi(t)$ 是原假设 $H_0: \theta=\theta_0$ 对备择假设 $H_1: \theta\neq\theta_0$ 的水平为 α 的 UMPUT. 定理证毕.

关于这个定理有三点说明.

1° 由定理 3.10 和 3.13 的证明过程知，单参数指数型分布族的双边假设检验问题(Ⅴ)的 UMPT 就是一致最优势边界相似检验；双边假设检验问题(Ⅳ)的 UMPUT 就是一致最优势边界相似检验. 而关于双边假设检验问题(Ⅲ)，由定理 3.14 的证明过程知，其 UMPUT 就是所有在边界上满足条件(3.37)和(3.38)的检验函数类中的一致最优势检验. 此外，如本节前面所述的，单参数指数型分布族的单边假设检验问题的 UMPT 就是一致最优势边界相似检验. 这些结论将被用于多参数指数型分布族的假设检验问题.

2° 如果在 $\theta=\theta_0$ 时，$T(x)$ 的分布关于某个数 r_0 是对称的，则双边假设检验问题(Ⅲ)的 UMPUT 可以简化为如下的形式，

$$\phi(T(x))=\begin{cases}1, & T(x)<r_0-\delta \text{ 或 } T(x)>r_0+\delta \\ r, & T(x)=r_0\pm\delta \\ 0, & r_0-\delta<T(x)<r_0+\delta\end{cases} \tag{3.41}$$

当其中的常数 δ 与 $r(0\leqslant r\leqslant 1)$ 由下式

$$P_{\theta_0}(T<r_0-\delta)+r\cdot P_{\theta_0}(T=r_0-\delta)=\frac{\alpha}{2}$$

确定时，(3.37)式与(3.38)式一定成立. 其证明如下：由 $T(x)$ 的分布关于 r_0 是对称的可以推得

$$\begin{aligned}E_{\theta_0}\phi(T) &=P_{\theta_0}(T<r_0-\delta)+P_{\theta_0}(T>r_0+\delta)+ \\ &\quad r\cdot P_{\theta_0}(T=r_0\pm\delta) \\ &=2\cdot\{P_{\theta_0}(T<r_0-\delta)+r\cdot P_{\theta_0}(T=r_0-\delta)\}=\alpha\end{aligned}$$

(3.37)式成立. 从而由 $E_{\theta_0}T=r_0$ 以及 $E_{\theta_0}[(T-r_0)\phi(T)]=0$ 可进一步推得

$$E_{\theta_0}[T\cdot\phi(T)]=E_{\theta_0}[(T-r_0)\phi(T)]+r_0\cdot E_{\theta_0}\phi(T)$$

$$= r_0 \cdot \alpha = \alpha \cdot \mathrm{E}_{\theta_0} T$$

(3.38)式也成立. 所以在 $\theta=\theta_0$ 时,如果 $T(x)$ 的分布关于某个数 r_0 是对称的,那么关于双边假设检验问题(Ⅲ)的 UMPUT 由(3.41)式给出.

3° 设 $T(x)$ 的分布密度函数形如 $p(t-\theta)$,其中 $p(t)$ 关于原点对称. 这说明 $T(x)$ 的分布关于 θ 是对称的,即对任意的 r 都有,$P_\theta(T<\theta-r)=P_\theta(T>\theta+r)$. 由 2° 知,关于双边假设检验问题(Ⅲ),只要取 $r_0=\theta_0$ 就可得到形如(3.41)的 UMPUT. 此外,关于双边假设检验问题(Ⅳ)与(Ⅴ),也有类似的结论. 对前者,取检验函数形如

$$\phi(T(x)) = \begin{cases} 1, & T(x) < \theta_1 - \delta \text{ 或 } T(x) > \theta_2 + \delta \\ r, & T(x) = \theta_1 - \delta \text{ 或 } T(x) = \theta_2 + \delta \\ 0, & \theta_1 - \delta < T(x) < \theta_2 + \delta \end{cases} \quad (3.42)$$

在 $T\sim p(t-\theta_1)$ 时,有 $T-\theta_1+\theta_2\sim p(t-\theta_2)$,并且对任意的 a,有 $P_\theta(T<\theta-a)=P_\theta(T>\theta+a)$. 所以

$$\begin{aligned} \mathrm{E}_{\theta_1}\phi(T) &= P_{\theta_1}(T < \theta_1 - \delta) + P_{\theta_1}(T > \theta_2 + \delta) + \\ &\quad r \cdot [P_{\theta_1}(T = \theta_1 - \delta) + P_{\theta}(T = \theta_2 + \delta)] \\ &= P_{\theta_2}(T < \theta_2 - \delta) + P_{\theta_2}(T > \theta_2 - \theta_2 - \theta_1 + \delta) + \\ &\quad r \cdot [P_{\theta_2}(T = \theta_2 - \delta) + P_{\theta_2}(T = \theta_2 + \theta_2 - \theta_1 + \delta)] \\ &= P_{\theta_2}(T > \theta_2 + \delta) + P_{\theta_2}(T < \theta_1 - \delta) + \\ &\quad r \cdot [P_{\theta_2}(T = \theta_2 + \delta) + P_{\theta_2}(T = \theta_1 - \delta)] \\ &= \mathrm{E}_{\theta_2}\phi(T) \end{aligned}$$

据定理 3.13 可知,如果(3.42)式中的常数 r 和 δ 由下式确定

$$\mathrm{E}_{\theta_1}\phi(T) = \alpha (\text{或 } \mathrm{E}_{\theta_2}\phi(T) = \alpha) \quad (3.43)$$

则由(3.42)式给出的检验函数是双边假设检验问题(Ⅳ)的 UMPUT. 类似地,取检验函数形如

$$\phi(T(x)) = \begin{cases} 1, & \theta_1 - \delta < T(x) < \theta_2 + \delta \\ r, & T(x) = \theta_1 - \delta \text{ 或 } T(x) = \theta_2 + \delta \\ 0, & T(x) < \theta_1 - \delta \text{ 或 } T(x) > \theta_2 + \delta \end{cases}$$

其中的常数 r 和 δ 由(3.43)式确定，则据定理 3.10 知，该检验函数是双边假设检验问题(Ⅴ)的 UMPT.

例 3.8 样本 $X=(X_1,\cdots,X_n)$ 来自正态分布族 $\{N(\mu,1):-\infty<\mu<\infty\}$. 考虑双边假设检验问题(1) $H_0:\mu=0$ 对 $H_1:\mu\neq 0$；(2) $H_0:-1\leqslant\mu\leqslant 1$ 对 $H_1:\mu<-1$ 或 $\mu>1$；(3) $H_0:\mu\leqslant -1$ 或 $\mu\geqslant 1$ 对 $H_1:-1<\mu<1$.

样本 X 的联合密度函数

$$p(x;\sigma^2)=c(\mu)\cdot\exp\{Q(\mu)\cdot T(x)\}\cdot h(x)$$

其中 $Q(\mu)=n\mu$ 是 μ 的严增函数，充分统计量 $T(X)=\overline{X}\sim N(\mu,1/n)$.

关于检验问题(1)，因为在 $\mu=0$ 时，$\overline{X}\sim N(0,1/n)$，其分布关于原点对称. 由上述 2° 的结论知，检验函数 $\phi(t)$ 的拒绝域取为 $\{x:\overline{x}\leqslant-\delta$ 或 $\overline{x}\geqslant\delta\}$，$\delta$ 由下式确定

$$E[\phi(T)\,|\,\mu=0]=1-P\{-\delta<\overline{X}<\delta\,|\,\mu=0\}=\alpha$$

则据定理 3.14 知，检验问题(1)的水平为 α 的 UMPUT 的拒绝域为 $\{x:|\overline{x}|\geqslant 1/\sqrt{n}\cdot U_{1-\alpha/2}\}$. $n=10$ 且 $\alpha=0.05$ 时，拒绝域为 $\{x:|\overline{x}|\geqslant 0.6198\}$. 其势函数的图形如图 3.6(1)所示.

关于检验问题(2)，因为充分统计量 $\overline{X}\sim N(\mu,1/n)$，其分布关于 μ 对称，则由上述 3° 的结论知，检验函数 $\phi(t)$ 的拒绝域取为 $\{x:\overline{x}\leqslant-1-\delta$ 或 $\overline{x}\geqslant 1+\delta\}$，$\delta$ 由下式确定

$$E[\phi(T)\,|\,\mu=1]=1-P\{-1-\delta<\overline{X}<1-\delta\,|\,\mu=1\}=\alpha$$

据定理 3.13 知，这是检验问题(2)的水平为 α 的 UMPUT. $n=10$ 且 $\alpha=0.05$ 时，δ 由下式确定

$$\begin{aligned}P\{-1-\delta<\overline{X}<1+\delta\,|\,\mu=1\}&=\Phi(\sqrt{10}\cdot\delta)-\Phi(\sqrt{10}(-2-\delta))\\&=0.95\end{aligned}$$

取 $\sqrt{10}\cdot\delta=U_{0.95}=1.64$，$\delta=0.5186$. 则有

$$\begin{aligned}P\{-1-\delta<\overline{X}<1+\delta\,|\,\mu=1\}&=\Phi(1.64)-\Phi(-7.9645)\\&\approx 0.95\end{aligned}$$

检验的拒绝域为 $\{x:|\overline{x}|\geqslant 1.5186\}$. 其势函数的图形如图 3.6(2)所示.

类似地，关于检验问题(3)，取检验函数 $\phi(t)$ 的拒绝域形如$\{x: -1-\delta \leqslant x \leqslant 1+\delta\}$. 若取 δ 满足条件

$$\mathrm{E}[\phi(T) \mid \mu = 1] = P\{-1-\delta \leqslant \overline{X} \leqslant 1+\delta\} = \alpha$$

则据定理 3.10 知，该检验为检验问题(3)的水平为 α 的 UMPT. $n=10$ 且 $\alpha=0.05$ 时，δ 由下式确定

$$P\{-1-\delta < \overline{X} < 1+\delta \mid \mu = 1\}$$
$$= \Phi(\sqrt{10} \cdot \delta) - \Phi(\sqrt{10}(-2-\delta)) = 0.05$$

取 $\sqrt{10} \cdot \delta = U_{0.05} = -1.64$，$\delta = -0.5186$. 则有

$$P\{-1-\delta < \overline{X} < 1+\delta \mid \mu = 1\} = \Phi(-1.64) - \Phi(-4.6846)$$
$$\approx 0.05$$

检验的拒绝域为$\{x: |\bar{x}| \leqslant 0.4814\}$. 其势函数的图形如图 3.6(3)所示.

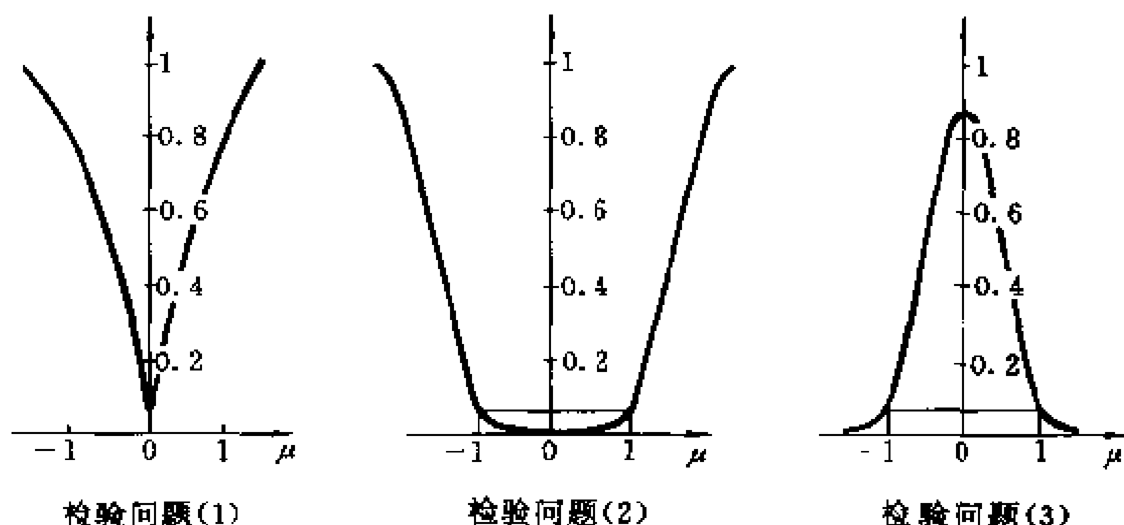


图 3.6　例 3.8 的势函数的图形

例 3.9　设样本 $X=(X_1,\cdots,X_n)$ 来自正态分布族$\{N(0,\sigma^2): \sigma^2>0\}$. 考虑检验问题

(1) $H_0: \sigma^2 \leqslant \sigma_1^2$ 或 $\sigma^2 \geqslant \sigma_2^2$, $H_1: \sigma_1^2 < \sigma^2 < \sigma_2^2$;

(2) $H_0: \sigma_1^2 \leqslant \sigma^2 \leqslant \sigma_2^2$, $H_1: \sigma^2 < \sigma_1^2$ 或 $\sigma^2 > \sigma_2^2$;

(3) $H_0: \sigma^2 = \sigma_0^2$, $H_1: \sigma^2 \neq \sigma_0^2$.

求问题(1)的水平为 $\alpha(0<\alpha<1)$的 UMPT；问题(2)和(3)的水平为 α $(0<\alpha<1)$的 UMPUT.

解:样本 X 的联合密度函数

$$p(x;\sigma^2)=c(\sigma^2)\cdot\exp\{Q(\sigma^2)\cdot T(x)\}$$

其中 $Q(\sigma^2)=(-2\sigma^2)^{-1}$是 σ^2 的严增函数，充分统计量 $T(X)=\sum_{i=1}^{n}X_i^2$ $\sim\sigma^2\cdot\chi^2(n)$. 这说明 $T(X)/\sigma^2\sim\chi^2(n)$.

根据定理 3.10，问题(1)的水平为 α 的 UMPT 的拒绝域为 $W=\{x:c_1\leqslant T(X)\leqslant c_2\}$，其中的 c_1 和 c_2 由下式确定

$$P_{\sigma_1^2}(c_1\leqslant T(X)\leqslant c_2)=P_{\sigma_2^2}(c_1\leqslant T(X)\leqslant c_2)=\alpha$$

这就是说，c_1 和 c_2 由下式确定

$$P_{\sigma_1^2}\left(\frac{c_1}{\sigma_1^2}\leqslant\frac{T(X)}{\sigma_1^2}\leqslant\frac{c_2}{\sigma_1^2}\right)=P_{\sigma_2^2}\left(\frac{c_1}{\sigma_2^2}\leqslant\frac{T(X)}{\sigma_2^2}\leqslant\frac{c_2}{\sigma_2^2}\right)=\alpha$$

因此 c_1 和 c_2 由以下两式确定

$$\int_{c_1/\sigma_1^2}^{c_2/\sigma_1^2}\chi^2(x|n)\mathrm{d}x=\alpha$$

$$\int_{c_1/\sigma_2^2}^{c_2/\sigma_2^2}\chi^2(x|n)\mathrm{d}x=\alpha$$

其中 $\chi^2(x|n)$表示自由度为 n 的 χ^2 分布的密度函数.

根据定理 3.13，问题(2)的水平为 α 的 UMPUT 的拒绝域为 $W=\{x:T(x)\leqslant c_1$ 或 $T(x)\geqslant c_2\}$. 与上述相类似地，其中的 c_1 和 c_2 由以下两式确定

$$\int_{c_1/\sigma_1^2}^{c_2/\sigma_1^2}\chi^2(x|n)\mathrm{d}x=1-\alpha$$

$$\int_{c_1/\sigma_2^2}^{c_2/\sigma_2^2}\chi^2(x|n)\mathrm{d}x=1-\alpha$$

根据定理 3.14，问题(3)的水平为 α 的 UMPUT$\phi(t)$的拒绝域为 $W=\{x:T(x)\leqslant c_1$ 或 $T(x)\geqslant c_2\}$. 由(3.37)式，则有

$$\mathrm{E}_{\sigma_0^2}[1-\phi(T)]=P_{\sigma_0^2}\left(\frac{c_1}{\sigma_0^2}\leqslant\frac{T(X)}{\sigma_0^2}\leqslant\frac{c_2}{\sigma_0^2}\right)=1-\alpha$$

即

$$\int_{c_1/\sigma_0^2}^{c_2/\sigma_0^2}\chi^2(x|n)\mathrm{d}x = 1-\alpha \tag{3.44}$$

由(3.38)式得 $\mathrm{E}_{\sigma_0^2}[T\phi(T)]=\alpha\cdot\mathrm{E}_{\sigma_0^2}T$. 由于 $\mathrm{E}_{\sigma_0^2}T=n\sigma_0^2$，从而有

$$\mathrm{E}_{\sigma_0^2}[T\cdot(1-\phi(T))]=n\sigma_0^2\cdot(1-\alpha)=n\sigma_0^2\cdot\mathrm{E}_{\sigma_0^2}[1-\phi(T)]$$

$$\mathrm{E}_{\sigma_0^2}\left[\left(\frac{T}{\sigma_0^2}-n\right)\cdot(1-\phi(T))\right]=0$$

这里$\dfrac{T}{\sigma_0^2}\sim\chi^2(n)$. 则有

$$\int_{c_1/\sigma_0^2}^{c_2/\sigma_0^2}(x-n)\cdot\chi^2(x|n)\mathrm{d}x=0 \tag{3.45}$$

故问题(3)的 UMPUT 中的 c_1 和 c_2 可由(3.44)和(3.45)两式确定. 另外，由势函数 $g(\sigma^2)=1-\int_{c_1/\sigma^2}^{c_2/\sigma^2}\chi^2(x|n)\mathrm{d}x$ 在 $\sigma^2=\sigma_0^2$ 时达到最小值，推得等式

$$c_2\cdot\chi^2\left(\frac{c_2}{\sigma_0^2}\Big|n\right)-c_1\cdot\chi^2\left(\frac{c_1}{\sigma_0^2}\Big|n\right)=0 \tag{3.46}$$

则问题(3)的 UMPUT 中，c_1 和 c_2 也可由(3.44)和(3.46)两式确定.

由(3.44)和(3.45)两式确定，或由(3.44)和(3.46)两式确定 c_1 和 c_2 过于复杂，不实用. 常简单地取

$$c_1=\chi^2_{\alpha/2}(n),c_2=\chi^2_{1-\alpha/2}(n)$$

这能使得(3.44)式成立，但不能使得(3.45)，或(3.46)式成立.

§3.5 多参数指数型分布族的假设检验

§3.5.1 多参数指数型分布族

单参数指数型分布族的单边和双边假设检验问题，基本上都被解决了。但是有些总体包含两个或两个以上的未知参数，当仅对其中一个未知参数做假设检验时，其它的未知参数就成了多余参数。一般来说，

即使对单边假设检验问题而言，其 UMPT 也是不存在的，但是有例外的情况。例如，样本 $X=(X_1,\cdots,X_n)$来自均值和方差皆未知的正态分布族$\{N(\mu,\sigma^2):-\infty<\mu<\infty,\sigma^2>0\}$。考虑关于方差 σ^2 的单边假设检验问题，(1)原假设 $H_0:\sigma^2\leqslant\sigma_0^2$ 对备择假设 $H_1:\sigma^2>\sigma_0^2$；(2)原假设 $H_0:\sigma^2\geqslant\sigma_0^2$对备择假设 $H_1:\sigma^2<\sigma_0^2$. 这时均值 μ 就是多余参数。常用的检验是 χ^2 检验。它们的拒绝域分别为$\left\{x:\sum(x_i-\bar{x})^2\geqslant\sigma_0^2\cdot\chi^2_{1-\alpha}(n-1)\right\}$和$\left\{x:\sum(x_i-\bar{x})^2\leqslant\sigma_0^2\cdot\chi^2_{\alpha}(n-1)\right\}$。Lehmann 和 Stein 在 1948 年证明了，问题(1)的 χ^2 检验是 UMPT，但问题(2)的 χ^2 检验不是 UMPT，问题(2)的 UMPT 是不存在的。本节将证明问题(2)的 χ^2 检验是 UMPUT。

一般来说，多参数指数型分布族的单边和双边假设检验问题，UMPT 是不存在的，但存在 UMPUT。对这类问题寻找 UMPUT 需要用到指数型分布族的有关性质。

设样本 X 服从多参数指数型分布，其密度函数为

$$p(x;\theta,\boldsymbol{r})=c(\theta,\boldsymbol{r})\cdot\exp\left\{\theta\cdot u(x)+\sum_{i=1}^{k}r_i\cdot t_i(x)\right\}\cdot h(x)\tag{3.47}$$

其中 θ 为被检验的参数，$\boldsymbol{r}=(r_1,\cdots,r_k)$为多余参数，$(\theta,\boldsymbol{r})\in\boldsymbol{\Omega}\subset\boldsymbol{R}^{k+1}$. 若参数空间 Ω 有内点，则$(U(X),T(X))=(U(X),T_1(X),\cdots,T_k(X))$是参数$(\theta,\boldsymbol{r})$的充分，完备的统计量. 据定理 3.1 的充分性原则，我们仅需考虑基于$(U(X),T(X))$的检验. $(U(X),T(X))$的分布仍为多参数指数型分布，其密度函数为

$$p(u,t;\theta,\boldsymbol{r})=c_1(\theta,\boldsymbol{r})\cdot\exp\left\{\theta\cdot u+\sum_{i=1}^{k}r_i\cdot t_i\right\}\cdot h_1(u,t)$$

在 $T(X)=t$ 的条件下，$U(X)$的条件分布与多余参数 $\boldsymbol{r}$ 无关，且为单参数指数型分布

$$p(u\,|\,T=t)=c_t(\theta)\cdot\exp\{\theta\cdot u\}\cdot h_t(u)\tag{3.48}$$

由此，多参数指数型分布族中的假设检验问题转化为单参数指数型分布族中的假设检验问题，即在$\{x:T(x)=t\}$的子空间上，寻找 UMPT，或 UMPUT. 这些在 $T(X)=t$ 的条件下最优的检验，是不是在整个样

本空间上一致最优的检验，这是我们还得讨论的问题.

$T(X)$的边际分布为

$$p(t) = c_\theta(\boldsymbol{r}) \cdot \exp\left\{\sum_{i=1}^{k} r_i \cdot t_i\right\} \cdot h_\theta(t)$$

在 θ 给定时，$T(X)$的边际分布为指数型分布. 令 $G_\theta = \{\boldsymbol{r}: (\theta, \boldsymbol{r}) \in \boldsymbol{\Omega}\}$. 在 θ 给定时，若空间 G_θ 有内点，则 $T(X)$是参数 $\boldsymbol{r}$ 的充分，完备的统计量. 由此我们可以证明，在 $T(X)=t$ 的条件下最优的检验，在整个样本空间上也是一致最优的检验.

§3.5.2 多参数指数型分布族的假设检验

考虑单边或双边假设检验问题(Ⅰ),…,(Ⅴ)(见§3.3.3). 设原假设 $H_0: \theta \in \Theta_0$，备择假设 $H_1: \theta \in \Theta_1$. 令 Γ 是 Θ_0 和 Θ_1 的公共边界. 则 Γ 仅含有一或两个点. 在 $\theta \in \Gamma$ 时，设 G_θ 有内点. 在 $T(X)=t$ 的条件下，基于条件单参数指数型分布(3.48)，若 $\phi(u,t)$是该检验问题的水平为 α 的(条件)UMPT，或(条件)UMPUT. 则可证明：在整个样本空间上，$\phi(u,t)$是该检验问题的水平为 α 的 UMPUT.

由于在 $T(X)=t$ 的条件下，$\phi(u,t)$是检验问题的水平为 α 的条件检验，所以

$$\mathrm{E}_\theta[\phi(U,T) \mid T = t] \leqslant \alpha, \quad \theta \in \Theta_0$$

从而

$$\mathrm{E}_{\theta,r}\phi(U,T) = \mathrm{E}_{\theta,r}\{\mathrm{E}_\theta[\phi(U,T) \mid T]\} \leqslant \alpha, \quad \theta \in \Theta_0$$

这说明 $\phi(u,t)$在整个样本空间上，也是该检验问题的水平为 α 的一个检验. 设 $\phi^*(u,t)$在整个样本空间上，是检验问题的水平为 α 的任意一个无偏检验. 对于指数族，任意一个检验的势函数都是连续的. 所以 $\mathrm{E}_{\theta,r}\phi^*(U,T)=\alpha, \theta \in \Gamma$. 从而在 $\theta \in \Gamma$ 给定时，对所有的 $\boldsymbol{r} \in G_\theta$，都有 $\mathrm{E}_{\theta,r}\{\mathrm{E}_\theta[\phi^*(U,T) \mid T]\} = \alpha$. 由于在 θ 给定时，$T(X)$是完备的统计量，因此在 $\theta \in \Gamma$ 给定时

$$\mathrm{E}_\theta[\phi^*(U,T) \mid T = t] = \alpha$$

这里可能有一个零测集 B，在 $t \in B$ 时，上述等式不成立. 但这并不影响后面的证明. 我们不妨假定上述等式总是成立的. 这说明在 $T(X)=t$

的条件下，基于条件单参数指数型分布(3.48)，$\phi^*(u,t)$是该检验问题的水平为 α 的边界相似检验.

首先考虑单边假设检验问题(Ⅰ)，(Ⅱ)和双边假设检验问题(Ⅳ)，(Ⅴ). 我们知道，单参数指数型分布族的这些检验问题的 UMPT，或 UMPUT 与一致最优势边界相似检验是等同的. 由于在 $T(X)=t$ 的条件下，基于条件单参数指数型分布(3.48)，$\phi(u,t)$是该检验问题的水平为 α 的条件 UMPT，或条件 UMPUT. 因此

$$\mathrm{E}_\theta[\phi(U,T)\,|\,T=t] \geqslant \mathrm{E}_\theta[\phi^*(U,T)\,|\,T=t], \quad \theta\in\boldsymbol{\Theta}_1$$

从而

$$\mathrm{E}_{\theta,r}\{\mathrm{E}_\theta[\phi(U,T)\,|\,T]\} \geqslant \mathrm{E}_{\theta,r}\{\mathrm{E}_\theta[\phi^*(U,T)\,|\,T]\}, \quad \theta\in\boldsymbol{\Theta}_1$$

$$\mathrm{E}_{\theta,r}\phi(U,T) \geqslant \mathrm{E}_{\theta,r}\phi^*(U,T), \quad \theta\in\boldsymbol{\Theta}_1$$

由于 $\phi^*(u,t)$在整个样本空间上，是水平为 α 的任意一个无偏检验，所以在整个样本空间上 $\phi(u,t)$是单边假设检验问题(Ⅰ)，(Ⅱ)和双边假设检验问题(Ⅳ)，(Ⅴ)的水平为 α 的 UMPUT. 至于在整个样本空间上 $\phi(u,t)$是双边假设检验问题(Ⅲ)的水平为 α 的 UMPUT 的证明，我们将它留作习题 3.12.

由上述结果易得下述定理.

定理 3.15 设样本 X 服从多参数指数型分布(3.47). 则

(1) 原假设 $H_0: \theta\leqslant\theta_0$ 对备择假设 $H_1: \theta>\theta_0$ 的单边假设检验问题(Ⅰ)的水平为 α 的 UMPUT 为

$$\phi(u,t)=\begin{cases}1, & u>c(t)\\ r(t), & u=c(t)\\ 0, & u<c(t)\end{cases} \tag{3.49}$$

其中 $r(t)(0\leqslant r(t)\leqslant 1)$和 $c(t)$由下式确定

$$\mathrm{E}_{\theta_0}[\phi(U,T)\,|\,T=t]=\alpha \tag{3.50}$$

如果考虑 $H_0: \theta\geqslant\theta_0$ 对 $H_1: \theta<\theta_0$ 的单边假设检验问题(Ⅱ)，类似的结论仍全部成立，但是(3.49)式中的不等号要改变方向.

(2) 原假设 $H_0: \theta_1\leqslant\theta\leqslant\theta_2$ 对备择假设 $H_1: \theta<\theta_1$ 或 $\theta>\theta_2$ 的双边假设检验问题(Ⅳ)的水平为 α 的 UMPUT 为

$$\phi(u,t)=\begin{cases}0, & c_1(t)<u<c_2(t)\\ r_i(t,) & u=c_i(t),i=1,2\\ 1, & u<c_1(t) \text{ 或 } u>c_2(t)\end{cases} \tag{3.51}$$

其中 $r_i(t)(0\leqslant r_i(t)\leqslant 1)$ 和 $c_i(t)$ 由下式确定

$$\mathrm{E}_{\theta_1}[\phi(U,T)|T=t]=\mathrm{E}_{\theta_2}[\phi(U,T)|T=t]=\alpha \tag{3.52}$$

如果考虑 $H_0:\theta\leqslant\theta_1$ 或 $\theta\geqslant\theta_2$ 对 $H_1:\theta_1<\theta<\theta_2$ 的双边假设检验问题(Ⅴ),类似的结论仍全部成立,但是(3.51)式中的"$c_1(t)<u<c_2(t)$"与"$u<c_1(t)$或 $u>c_2(t)$"要互相交换.

(3)原假设 $H_0:\theta=\theta_0$ 对备择假设 $H_1:\theta\neq\theta_0$ 的双边假设检验问题(Ⅲ)的水平为 α 的 UMPUT 由(3.51)式给出,其中 $r_i(t)(0\leqslant r_i(t)\leqslant 1)$和 $c_i(t)$由(3.50)和以下的式子确定

$$\mathrm{E}_{\theta_0}[U\phi(U,T)|T=t]=\alpha\cdot\mathrm{E}_{\theta_0}[U|T=t] \tag{3.53}$$

如果所要检验的参数是 $\theta^*=a\theta+\sum_{i=1}^k b_i r_i(a\neq 0)$,只需将(3.47)式改写成下面的形式

$$p(x;\theta,\boldsymbol{r})=c(\theta,\boldsymbol{r})\cdot\exp\left\{\theta^*\cdot u^*(x)+\sum_{i=1}^k r_i\cdot t_i^*(x)\right\}\cdot h(x)$$

其中 $u^*(x)=\dfrac{u(x)}{a}$,$t_i^*(x)=t_i(x)-\dfrac{b_i}{a}u(x)$,利用定理 3.15 就可解决关于 θ^* 的各种单边和双边假设检验问题.

§ 3.5.3　两个 Poisson 总体的比较

设相互独立的两个总体 x 和 y 的分布分别是参数为 λ 和 μ 的 Poisson 分布.要比较这两个总体的差异,就是要比较参数 λ 和 μ 的差异.为了应用定理 3.15,首先把它们的联合分布表示成指数族分布(3.47)的形式

$$P(x,y;\lambda,\mu)=\mathrm{e}^{-(\lambda+\mu)}\cdot\exp\{\theta\cdot u+r\cdot t\}\cdot(x!y!)^{-1}$$

其中 $\theta=\ln(\mu/\lambda)$,$r=\ln(\lambda)$,$u=y$,$t=x+y$.比较参数 λ 和 μ 的差异的检验问题就变为关于参数 θ 的检验问题.例如原假设 $H_0:\lambda\leqslant\mu$ 对备择假设 $H_1:\lambda>\mu$ 的检验问题等价于原假设 $H_0:\theta\geqslant 0$ 对备择假设 $H_1:\theta<0$ 的检验问题.按定理 3.15,只要在直线 $x-y=t$ 的整数点上寻

找最优势的条件检验. 在给定 $X+Y=t$ 后, Y 的条件分布为

$$P(y|x+y=t)=\binom{t}{y}\cdot p^y\cdot(1-p)^{t-y}, y=0,1,\cdots,t$$

这是二项分布 $b(t,p)$. 其中 $p=\frac{\mu}{\lambda+\mu}=\frac{e^\theta}{1+e^\theta}$. 所以原来的假设检验问题就化为关于二项分布的参数 p 的假设检验问题. 例如原假设 H_0 : $\lambda\leqslant\mu$ 对备择假设 H_1 : $\lambda>\mu$ 的检验问题等价于原假设 H_0 : $p\geqslant0.5$ 对备择假设 H_1 : $p<0.5$ 假设检验问题.

上述方法也可用来检验两个样本 $X_1,\cdots,X_m$ 和 $Y_1,\cdots,Y_n$ 是否来自同一个 Poisson 总体. 这是因为 $\sum_{i=1}^m X_i$ 和 $\sum_{i=1}^n Y_i$ 分别是参数为 $m\lambda$ 和 $n\mu$ 的 Poisson 分布. 检验这两个样本是否来自同一个 Poisson 总体, 等价于检验 $m\lambda$ 和 $n\mu$ 的比值是否等于 $m:n$.

§3.5.4 两个二项总体的比较

设相互独立的两个总体 x 和 y 皆服从二项分布. 它们分别是 $b(m,p_1)$ 和 $b(n,p_2)$. 它们的联合分布可表示成指数族分布(3.47)的形式

$$P(x,y;p_1,p_2)=(1-p_1)^m\cdot(1-p_2)^n\cdot \exp\{\theta\cdot u+r\cdot t\}\cdot\binom{x}{m}\cdot\binom{y}{n}$$

其中 $\theta=\ln\frac{p_2(1-p_1)}{p_1(1-p_2)}$, $r=\ln\frac{p_1}{1-p_1}$, $u=y$, $t=x+y$. 那么比较这两个总体的差异, 即比较参数 p_1 和 p_2 的差异的检验问题就变为关于参数 θ 的检验问题. 例如原假设 H_0 : $p_1=p_2$ 对备择假设 H_1 : $p_1\neq p_2$ 的检验问题等价于原假设 H_0 : $\theta=0$ 对备择假设 H_1 : $\theta\neq0$ 的检验问题. 按定理 3.15, 只要在直线 $x+y=t$ 的整数点上寻找最优势的条件检验. 在给定 $x+y=t$ 后, y 的条件分布为

$$P(y|x+y=t)=c_t(\theta)\cdot\binom{m}{t-y}\binom{n}{y}\cdot e^{\theta y},\quad y=0,1,\cdots,t$$

其中 $c_t(\theta)=\left[\sum_{y=0}^{t}\binom{m}{t-y}\binom{n}{y}\cdot e^{\theta y}\right]^{-1}$. 在 $\theta=0$ 时, 这个条件分布是

超几何分布

$$P(y|x+y=t)=\binom{m}{t-y}\binom{n}{y}\Big/\binom{m+n}{t},\quad y=0,1,\cdots,t$$

用它可以确定原假设 $H_0: \theta=0$ 对备择假设 $H_1: \theta\neq 0$ 的检验问题的临界值.

§ 3.5.5 正态总体参数的检验问题

为了能用更简单的形式给出正态总体参数的检验问题的检验函数,首先证明下述定理.

定理 3.16 设样本 X 服从多参数指数型分布(3.47).

(1)如果存在一个统计量 $V=V(U,T)$,给定 $T=t$ 后,V 是 U 的单调增函数,且当 $\theta=\theta_0$ 时,V 与 T 互相独立. 则 $H_0: \theta\leqslant\theta_0$ 对 $H_1: \theta>\theta_0$ 的单边假设检验问题(Ⅰ)的水平为 α 的 UMPUT 为

$$\phi(v)=\begin{cases}1, & v>c\\ r, & v=c\\ 0, & v<c\end{cases}\tag{3.54}$$

其中 $r(0\leqslant r\leqslant 1)$和 c 不依赖于 t. 它们由下式确定

$$\mathrm{E}_{\theta_0}\phi(V)=\alpha\tag{3.55}$$

如果考虑 $H_0: \theta\geqslant\theta_0$ 对 $H_1: \theta<\theta_0$ 的单边假设检验问题(Ⅱ),类似的结论仍全部成立,但是(3.54)式中的不等号要改变方向.

(2) 如果存在一个统计量 $V=V(U,T)$,给定 $T=t$ 后,V 是 U 的单调函数,且当 $\theta=\theta_1$ 或 θ_2 时,V 与 T 互相独立. 则 $H_0: \theta_1\leqslant\theta\leqslant\theta_2$ 对 $H_1: \theta<\theta_1$ 或 $\theta>\theta_2$ 的双边假设检验问题(Ⅳ)的水平为 α 的 UMPUT 为

$$\phi(v)=\begin{cases}0, & c_1<v<c_2\\ r_i, & v=c_i, i=1,2\\ 1, & v<c_1 \text{ 或 } v>c_2\end{cases}\tag{3.56}$$

其中 $r_i(0\leqslant r_i\leqslant 1)$和 c_i 不依赖于 t. 它们由下式确定

$$\mathrm{E}_{\theta_1}\phi(V)=\mathrm{E}_{\theta_2}\phi(V)=\alpha\tag{3.57}$$

如果考虑 $H_0: \theta\leqslant\theta_1$ 或 $\theta\geqslant\theta_2$ 对 $H_1: \theta_1<\theta<\theta_2$ 的双边假设检验问题

(Ⅴ),类似的结论仍全部成立,但是(3.56)式中的"$c_1<v<c_2$"与"$v<c_1$ 或 $v>c_2$"要互相交换.

(3) 如果存在一个统计量 $V=a(T)\cdot U+b(T)(a(T)\neq 0)$,即给定 $T=t$ 后,V 是 U 的线性函数,且当 $\theta=\theta_0$ 时,V 与 T 互相独立.则 $H_0:\theta=\theta_0$ 对 $H_1:\theta\neq\theta_0$ 的双边假设检验问题(Ⅲ)的水平为 α 的 UMPUT 由(3.56)式给出,其中的 $r_i(0\leqslant r_i\leqslant 1)$和 c_i 不依赖于 t. 它们由(3.55)和以下的式子确定

$$\mathrm{E}_{\theta_0}[V\cdot\phi(V)]=\alpha\cdot\mathrm{E}_{\theta_0}V \tag{3.58}$$

证明:(1)据定理 3.15(1),单边假设检验问题(Ⅰ)的水平为 α 的 UMPUT 由(3.49)和(3.50)两式给出.因为对给定的 t,$v=v(u,t)$是 u 的单调增函数,所以检验函数等价地由下式给出

$$\phi(v,t)=\begin{cases}1, & v>c(t)\\ r(t), & v=c(t)\\ 0, & v<c(t)\end{cases} \tag{3.59}$$

其中 $r(t)(0\leqslant r(t)\leqslant 1)$和 $c(t)$由 $\mathrm{E}_{\theta_0}[\phi(V,T)\,|\,T=t]=\alpha$ 确定.又因为在 $\theta=\theta_0$ 时,V 与 T 互相独立,所以 $r(t)$和 $c(t)$不依赖于 t,而由(3.55)式给出.则由(3.59)式给出的检验就是由(3.54)式给出的检验.关于单边假设检验问题(Ⅱ)有类似的结论.(1)得证.

(2)据定理 3.15(2),双边假设检验问题(Ⅳ)的水平为 α 的 UMPUT 由(3.51)和(3.52)两式给出.因为对给定的 t,$v=v(u,t)$是 u 的单调函数,又因为在 $\theta=\theta_1$ 或 θ_2 时,V 与 T 互相独立,所以检验函数等价地由(3.56)式给出,且 $r_i(0\leqslant r_i\leqslant 1)$和 c_i 不依赖于 t,而由(3.57)式给出.关于双边假设检验问题(Ⅴ)有类似的结论.(2)得证.

(3)据定理 3.15(3),双边假设检验问题(Ⅲ)的水平为 α 的 UMPUT 由(3.51),(3.50)和(3.53)三式给出.因为对给定的 t,$v=a(t)u+b(t)$,v 是 u 的单调函数,所以检验函数等价地由下式给出

$$\phi(v,t)=\begin{cases}0, & c_1(t)<v<c_2(t)\\ r_i(t), & v=c_i(t),i=1,2\\ 1, & v<c_1(t)\text{ 或 }v>c_2(t)\end{cases} \tag{3.60}$$

其中 $r(t)_i(0\leqslant r_i(t)\leqslant 1)$和 $c_i(t)$由以下两式确定,

$$E_{\theta_0}[\phi(V,T)|T=t]=\alpha$$

$$E_{\theta_0}\left[\frac{V-b(T)}{a(T)}\phi(V,T)|T=t\right]=\alpha\cdot E_{\theta_0}\left[\frac{V-b(T)}{a(T)}|T=t\right]$$

由此可以推得，它们由以下两式确定，

$$E_{\theta_0}[\phi(V,T)|T=t]=\alpha$$

$$E_{\theta_0}[V\cdot\phi(V,T)|T=t]=\alpha\cdot E_{\theta_0}[V|T=t]$$

因为在 $\theta=\theta_0$ 时，V 与 T 互相独立，所以 $r_i(t)$ $(0\leqslant r_i(t)\leqslant 1)$ 和 $c_i(t)$ 不依赖于 t，而由(3.55)和(3.58)式确定. 则由(3.60)式给出的检验就是由(3.56)式给出的检验. (3)得证. 定理证毕.

为了能用更简单的形式给出正态总体参数的检验问题的检验函数，尚需下述引理.

引理 3.17 设 $(\mathscr{X},\mathscr{B},\mathscr{P}=\{P_\theta:\theta\in\Theta\})$ 为一参数统计结构，$T(X)$ 为 θ 的充分且完备的统计量. 如果可测统计量 $S(X)$ 的分布与 θ 无关，则对所有的 $\theta\in\Theta$，$T(X)$ 与 $S(X)$ 相互独立.

证明见参考文献[2].

例 3.10 设样本 $X=(X_1,\cdots,X_n)$ 来自正态总体 $N(\mu,\sigma^2)$，其中 $-\infty<\mu<\infty$，$\sigma^2>0$. 考虑关于 μ 的假设检验问题，(1) $H_0:\mu\geqslant 0$ 对 $H_1:\mu<0$；(2) $H_0:\mu\leqslant 0$ 对 $H_1:\mu>0$；(3) $H_0:\mu=0$ 对 $H_1:\mu\neq 0$. 若考虑的假设检验问题是 $H_0:\mu\geqslant\mu_0$ 对 $H_1:\mu<\mu_0$ 等，则可以通过变换 $Y_i=X_i-\mu_0$，归结为上述情况之一.

样本的联合密度为

$$p(x;\mu,\sigma^2)=(\sqrt{2\pi}\sigma)^{-n}\exp\left\{-\frac{n\mu^2}{2\sigma^2}\right\}\cdot\exp\{\theta\cdot u+r\cdot t\}$$

其中 $\theta=n\mu/\sigma^2$，$r=(-2\sigma^2)^{-1}$，$u=\bar{x}$，$t=\sum_{i=1}^n x_i^2$. 这时，关于 μ 的检验问题就变为关于 θ 的检验问题. 例如，检验问题(1)等价于 $H_0:\theta\geqslant 0$ 对 $H_1:\theta<0$ 的检验问题.

为了能用定理 3.16(1)，以较为简单的形式给出问题(1)和(2)的检验函数，首先寻找统计量 $V=V(U,T)$，使得对给定的 T，V 是 U 的单调增函数，且在 $\theta=0$，即在 $\mu=0$ 时，V 与 T 互相独立. 为此令

$$V=\sqrt{n(n-1)}\cdot\frac{U}{\sqrt{T-nU^2}}=\sqrt{n(n-1)}\cdot\frac{\overline{X}}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}}$$

下面证明:它满足上述要求. 对给定的 T,V 是 U 的单调增函数. 在 $\mu=0$ 时,样本 $X=(X_1,\cdots,X_n)$来自正态总体 $N(0,\sigma^2)$. 这时, $T=\sum_{i=1}^{n}X_i^2$ 是 σ^2 的充分、完备的统计量. 且在 $\mu=0$ 时,由于

$$V=\sqrt{n(n-1)}\cdot\frac{\overline{X}}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}}=\sqrt{n(n-1)}\cdot\frac{\overline{Z}}{\sqrt{\sum_{i=1}^{n}(Z_i-\overline{Z})^2}}$$

其中 $Z_i=X_i/\sigma\sim N(0,1)$,所以 V 的分布与参数 σ^2 无关. 因为在 $\mu=0$ 时,T 是充分、完备的统计量,而 V 的分布又与参数无关,则据引理 3.17知,V 与 T 互相独立. 在 $\mu=0$ 时,$V\sim t(n-1)$. 根据定理 3.16(1),前两个检验问题的 UMPUT 的拒绝域分别为$\{x: v\leqslant t_\alpha(n-1)\}$和$\{x: v\geqslant t_{1-\alpha}(n-1)\}$. 这就是常用的 t 检验法.

为了能用定理 3.16(3),以较为简单的形式给出问题(3)的检验函数,首先寻找统计量 W. 它形如 $W=a(T)U+b(T)$ $(a(T)\neq0)$,且在 $\theta=0$ 时,W 与 T 互相独立. 为此令

$$W=\frac{U}{\sqrt{T}}=\frac{\overline{X}}{\sqrt{\sum_{i=1}^{n}X_i^2}}$$

则对固定的 T,W 是 U 的线性函数;且在 $\mu=0$ 时,由于 $T=\sum_{i=1}^{n}X_i^2$ 是 σ^2 的充分、完备的统计量,且由于 W 的分布与参数 σ^2 无关(习题 3.13),所以据引理 3.17 知,W 与 T 互相独立. 根据定理 3.16(3),并由于在 $\mu=0$ 时,W 的分布关于原点对称(习题 3.13),从而根据定理 3.14的说明,最后一个检验问题的 UMPUT 的拒绝域为$\{x: |w|\geqslant c\}$. 由于

$$V=\sqrt{n(n-1)}\frac{W}{\sqrt{1-nW^2}}$$

故$|V|$是$|W|$的单调增函数. 从而拒绝域可等价地表示为$\{x: |v|\geqslant$

$c\}$. 由于在 $\mu=0$ 时, $V\sim t(n-1)$. 从而该检验问题的 UMPUT 的拒绝域为 $\{x: |v|\geqslant t_{1-\alpha/2}(n-1)\}$. 这也是常用的 t 检验法.

例 3.11 设样本 $X=(X_1,\cdots,X_n)$ 来自正态总体 $N(\mu,\sigma^2)$, 其中 $-\infty<\mu<\infty$, $\sigma^2>0$. 考虑关于 σ^2 的假设检验问题. (1) $H_0: \sigma^2\geqslant\sigma_0^2$ 对 $H_1: \sigma^2<\sigma_0^2$; (2) $H_0: \sigma^2\leqslant\sigma_0^2$ 对 $H_1: \sigma^2>\sigma_0^2$; (3) $H_0: \sigma^2=\sigma_0^2$ 对 $H_1: \sigma^2\neq\sigma_0^2$. 样本的联合密度为

$$p(x;\mu,\sigma^2)=(\sqrt{2\pi}\sigma)^{-n}\exp\left\{-\frac{n\mu^2}{2\sigma^2}\right\}\cdot\exp\{\theta\cdot u+r\cdot t\}$$

其中 $\theta=(-2\sigma^2)^{-1}$, $r=n\mu/\sigma^2$, $u=\sum_{i=1}^n x_i^2$, $t=\bar{x}$. 这时, 关于 σ^2 的检验问题就变为关于 θ 的检验问题, 例如检验问题(1)就变为 $H_0: \theta\geqslant(-2\sigma_0^2)^{-1}$ 对 $H_1: \theta<(-2\sigma_0^2)^{-1}$.

令 $V=U-n\cdot T^2=\sum_{i=1}^n(X_i-\overline{X})^2$. 则对固定的 $T=\overline{X}$, V 是 U 的线性增函数. 显然, V 与 T 相互独立. 在 $\theta=(-2\sigma_0^2)^{-1}$, 即 $\sigma^2=\sigma_0^2$ 时, $V\sim\sigma_0^2\cdot\chi^2(n-1)$. 根据定理 3.16(1), 前两个检验问题的 UMPUT 的拒绝域分别为 $\{x: v\leqslant\sigma_0^2\cdot\chi_\alpha^2(n-1)\}$ 和 $\{x: v\geqslant\sigma_0^2\cdot\chi_{1-\alpha}^2(n-1)\}$. 这就是常用的 χ^2 检验法. 最后一个检验问题的 UMPUT 的拒绝域为 $\{x: v\leqslant c_1$ 或 $x: v\geqslant c_2\}$, 其中 c_1 与 c_2 由(3.44)与(3.45), 或(3.44)与(3.46)两式确定, 但需把其中的自由度 n 换为 $n-1$. 一般常取 $c_1=\chi_{\alpha/2}^2(n-1)$, $c_2=\chi_{1-\alpha/2}^2(n-1)$. 这也是常用的 χ^2 检验法.

例 3.12 设样本 $X=(X_1,\cdots,X_m)$ 和样本 $Y=(Y_1,\cdots,Y_n)$ 互相独立. 它们分别来自正态总体 $N(\mu_1,\sigma^2)$ 和 $N(\mu_2,\sigma^2)$, 其中: $-\infty<\mu_1<\infty$, $-\infty<\mu_2<\infty$, $\sigma^2>0$. 由于这两个正态总体的方差相等, 所以比较这两个正态总体的差异的检验问题就是比较这两个正态总体的均值的检验问题, 例如 $H_0: \mu_1=\mu_2$ 对 $H_1: \mu_1\neq\mu_2$.

样本的联合密度为

$$p(x,y;\mu_1,\mu_2,\sigma^2)=(2\pi\sigma^2)^{-n}\exp\left\{-\frac{n(\mu_1^2+\mu_2^2)}{2\sigma^2}\right\}\cdot$$
$$\exp\{\theta\cdot u+r_1\cdot t_1+r_2\cdot t_2\}$$

其中

$$\theta=\frac{mn(\mu_2-\mu_1)}{(m+n)\sigma^2},r_1=\frac{m\mu_1+n\mu_2}{(m+n)\sigma^2},r_2=-\frac{1}{2\sigma^2}$$

$$u=\bar{y}-\bar{x},t_1=m\bar{x}+n\bar{y},t_2=\sum_{i=1}^{m}x_i^2+\sum_{i=1}^{n}y_i^2$$

这时，所考虑的检验问题 $H_0:\mu_1=\mu_2$ 对 $H_1:\mu_1\neq\mu_2$ 就变为关于 θ 的检验问题，$H_0:\theta=0$ 对 $H_1:\theta\neq 0$.

根据定理 3.16(3)，首先寻找统计量 V，使得对固定的 T_1 与 T_2，V 是 U 的线性函数，且在 $\theta=0$ 时，V 与 (T_1,T_2) 互相独立. 为此令

$$V=\frac{U}{\sqrt{T_2-\frac{T_1^2}{m+n}}}$$

则对固定的 T_1 与 T_2，V 是 U 的线性函数. 在 $\theta=0$，即在 $\mu_1=\mu_2=\mu$ 时，合样本 $(X_1,\cdots,X_m,Y_1,\cdots,Y_n)$ 不仅互相独立，而且同分布. 这时，σ^2 和 μ 的充分、完备的统计量为 T_1 和 T_2. 由于在 $\theta=0$ 时，V 的分布与参数 σ^2 和 μ 无关(习题 3.14)，所以据引理 3.17 知，V 与 (T_1,T_2) 互相独立. 根据定理 3.16(3)，并由于在 $\theta=0$ 时，V 的分布关于原点对称(习题 3.14)，从而根据定理 3.14 的说明，该检验问题的 UMPUT 的拒绝域为 $\{x,y:|v|\geqslant c\}$. 进一步令

$$W=\sqrt{\frac{mn(m+n-2)}{m+n}}\frac{V}{\sqrt{1-\frac{mn}{m+n}V^2}}$$

$$=\sqrt{\frac{mn(m+n-2)}{m+n}}\cdot\frac{\bar{Y}-\bar{X}}{\sqrt{\sum_{i=1}^{m}(X_i-\bar{X})^2+\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

则 $|V|$ 是 $|W|$ 的单调增函数，并且在 $\theta=0$ 时，$W\sim t(m+n-2)$. 所以该检验问题的 UMPUT 的拒绝域为 $\{x,y:|w|\geqslant t_{1-\alpha/2}(m+n-2)\}$. 这就是常用的 t 检验法. 类似地，检验问题(1)$H_0:\mu_1\leqslant\mu_2$ 对 $H_1:\mu_1>\mu_2$ 和(2)$H_0:\mu_1\geqslant\mu_2$ 对 $H_1:\mu_1<\mu_2$ 的 UMPUT 的拒绝域分别为 $\{x,y:w\leqslant t_{\alpha/2}(m+n-2)\}$ 和 $\{x,y:w\geqslant t_{1-\alpha/2}(m+n-2)\}$. 它们都是常用的 t 检验法.

例 3.13 设样本 $X=(X_1,\cdots,X_m)$ 和样本 $Y=(Y_1,\cdots,Y_n)$ 互相独立. 它们分别来自正态总体 $N(\mu_1,\sigma_1^2)$ 和 $N(\mu_2,\sigma_2^2)$, 其中 $-\infty<\mu_1<\infty$, $-\infty<\mu_2<\infty$, $\sigma_1^2>0$, $\sigma_2^2>0$. 考虑比较这两个正态总体的方差的检验问题, $H_0:\sigma_1^2=\sigma_2^2$ 对 $H_1:\sigma_1^2\neq\sigma_2^2$.

样本的联合密度为

$$p(x,y;\mu_1,\mu_2,\sigma_1^2,\sigma_2^2)=c(\mu_1,\mu_2,\sigma_1,\sigma_2)\cdot \exp\{\theta\cdot u+r_1\cdot t_1+r_2\cdot t_2+r_3\cdot t_3\}$$

其中

$$\theta=\frac{1}{2\sigma_1^2}-\frac{1}{2\sigma_2^2},r_1=-\frac{1}{2\sigma_1^2},r_2=\frac{m\mu_1}{\sigma_1^2},r_3=\frac{n\mu_2}{\sigma_2^2}$$

$$u=\sum_{i=1}^n y_i^2,t_1=\sum_{i=1}^m x_i^2+\sum_{i=1}^n y_i^2,t_2=\bar{x},t_3=\bar{y}$$

这时, 关于 σ^2 的检验问题 $H_0:\sigma_1^2=\sigma_2^2$ 对 $H_1:\sigma_1^2\neq\sigma_2^2$, 就变为关于 θ 的检验问题 $H_0:\theta=0$ 对 $H_1:\theta\neq0$.

根据定理 3.16(3), 首先寻找统计量 V, 使得对固定的 T_1,T_2 与 T_3, V 是 U 的线性函数, 且在 $\theta=0$ 时, V 与 (T_1,T_2,T_3) 互相独立. 为此令

$$V=\frac{U-nT_3^2}{T_1-mT_2^2-nT_3^2}=\frac{\sum_{i=1}^n(Y_i-\bar{Y})^2}{\sum_{i=1}^m(X_i-\bar{X})^2+\sum_{i=1}^n(Y_i-\bar{Y})^2}$$

则对给定的 T_1,T_2 和 T_3, V 是 U 的线性增函数. 在 $\theta=0$, 即在 $\sigma_1^2=\sigma_2^2=\sigma^2$ 时, μ_1,μ_2 和 σ^2 的充分、完备的统计量为

$$T_2=\bar{X},T_3=\bar{Y} \text{ 和 } T_1-mT_2^2-nT_3^2$$
$$=\sum_{i=1}^m(X_i-\bar{X})^2+\sum_{i=1}^n(Y_i-\bar{Y})^2$$

由于在 $\theta=0$ 时, V 的分布与参数 μ_1,μ_2 和 σ^2 无关(习题 3.15), 所以据引理 3.17 知, V 与 $(T_2,T_3,T_1-mT_2^2-nT_3^2)$, 从而与 (T_1,T_2,T_3) 互相独立. 由于在 $\theta=0$ 时, $V\sim Be\left(\frac{n-1}{2},\frac{m-1}{2}\right)$, 根据定理 3.16(3), 检验问题 $H_0:\sigma_1^2=\sigma_2^2$ 对 $H_1:\sigma_1^2\neq\sigma_2^2$ 的 UMPUT 的拒绝域为 $\{x,y:v\leqslant c_1$

或 $v\geqslant c_2\}$,其中 c_1 和 c_2 由以下两式确定(习题 3.15)

$$\int_{c_1}^{c_2} Be\left(x\left|\frac{n-1}{2},\frac{m-1}{2}\right.\right)\mathrm{d}x = 1-\alpha$$

$$\int_{c_1}^{c_2}\left(x-\frac{n-1}{m+n-2}\right)\cdot Be\left(x\left|\frac{n-1}{2},\frac{m-1}{2}\right.\right)\mathrm{d}x = 0$$

这里 $Be\left(x\left|\frac{n-1}{2},\frac{m-1}{2}\right.\right)$ 表示 $Be\left(\frac{n-1}{2},\frac{m-1}{2}\right)$ 分布的密度函数. 令

$$F=\frac{V}{1-V}\cdot\frac{m-1}{n-1}=\frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}{\sum_{i=1}^{m}(X_i-\overline{X})^2}\cdot\frac{m-1}{n-1}$$

则在 $\theta=0$ 时,$F\sim F(n-1,m-1)$. 由于 F 是 V 的严增函数,所以该检验问题的 UMPUT 的拒绝域可等价地写为 $\{x,y: F\leqslant c_1$ 或 $F\geqslant c_2\}$,其中 c_1 和 c_2 由以下两式确定(习题 3.15)

$$\int_{c_1}^{c_2} F(x|n-1,m-1)\mathrm{d}x = 1-\alpha$$

$$\int_{c_1}^{c_2}\frac{x-1}{(n-1)x+(m-1)}\cdot F(x|n-1,m-1)\mathrm{d}x = 0$$

这里 $F(x|n-1,m-1)$ 表示 $F(n-1,m-1)$ 分布的密度函数.

拒绝域为 $\{x,y: F\leqslant c_1$ 或 $F\geqslant c_2\}$ 的检验的势函数与参数 μ_1 和 μ_2 无关,仅依赖于比值 $\rho=\frac{\sigma_1^2}{\sigma_2^2}$. 它的势函数为

$$g(\rho)=1-\int_{\rho\cdot c_1}^{\rho\cdot c_2} F(x|n-1,m-1)\mathrm{d}\mu(x)$$

与例 3.9 相类似地,由势函数在 $\rho=1$ 时达到最小值推得,c_1 与 c_2 也可由以下两式确定,

$$\int_{c_1}^{c_2} F(x|n-1,m-1)\mathrm{d}x = 1-\alpha$$

$$c_2\cdot F(c_2|n-1,m-1)=c_1\cdot F(c_1|n-1,m-1)$$

一般常取 $c_1=F_{\alpha/2}(n-1,m-1)$,$c_2=F_{1-\alpha/2}(n-1,m-1)$. 这是常用的 F 检验法.

由于 F 可改写为

$$F = \frac{U - nT_3^2}{T_1 - mT_2^2 - U} \cdot \frac{m-1}{n-1}$$

则对给定的 T_1, T_2 和 T_3，F 是 U 的严增函数. 前已证明，在 $\theta=0$ 时，V 与 (T_1, T_2, T_3) 互相独立，而 F 仅与 V 有关，所以 F 与 (T_1, T_2, T_3) 也互相独立. 根据定理 3.16(1)，检验问题 $H_0: \sigma_1^2 \leqslant \sigma_2^2$ 对 $H_1: \sigma_1^2 > \sigma_2^2$，以及 $H_0: \sigma_1^2 \geqslant \sigma_2^2$ 对 $H_1: \sigma_1^2 < \sigma_2^2$ 的 UMPUT 的拒绝域分别为 $\{x, y: F \leqslant F_\alpha(n-1, m-1)\}$ 和 $\{x, y: F \geqslant F_{1-\alpha}(n-1, m-1)\}$. 这些都是常用的 F 检验法.

§3.6 似然比检验

§3.6.1 似然比检验

前面讨论了一致最优势检验和一致最优势无偏检验。但是在许多情况下，对于一个复合假设的检验问题，上述这些最优势检验可能不存在。这时，我们当然可以采用类似于前面的处理方法，对检验提出进一步的限制，然后在这被限制的较小的检验类中寻找最优检验，这样做，在实际使用中往往是困难的，也是不可取的。另一个途径是给出一个构造检验法的一般方法。N-P 基本引理告诉我们，在简单原假设对简单备择假设的检验问题中，最优势检验由似然比检验给出。事实上，似然比检验方法也可用在复合假设的检验问题中构造检验。这是构造检验的常用方法。

设 $X=(X_1, \cdots, X_n)$ 的分布密度函数是 $p(x;\theta)$，其中未知参数 $\theta \in \boldsymbol{\Theta}$ 可以是向量。我们知道，简单原假设 $H_0: \theta=\theta_0$ 对简单备择假设 $H_1: \theta=\theta_1 (\theta_1 \neq \theta_0)$ 的检验问题的似然比为

$$\lambda(x) = \frac{p(x;\theta_1)}{p(x;\theta_0)}$$

现考虑复合假设的检验问题：原假设 $H_0: \theta \in \Theta_0$ 对备择假设 $H_1: \theta \in \Theta_1$. 在复合假设的检验问题中，很自然地定义似然比 $\lambda(x)$ 为

$$\lambda(x)=\frac{\sup\limits_{\theta\in\Theta_1}p(x;\theta)}{\sup\limits_{\theta\in\Theta_0}p(x;\theta)}=\frac{p(x;\hat{\theta}_1)}{p(x;\hat{\theta}_0)}$$

其中，$\hat{\theta}_0$ 和 $\hat{\theta}_1$ 分别是 H_0 和 H_1 成立时，θ 的 MLE. $p(x;\hat{\theta}_0)$是原假设 H_0 成立时，观察到样本点 x 的可能性的一个度量，而 $p(x;\hat{\theta}_1)$是备择假设 H_1 成立时，观察到样本点 x 的可能性的一个度量. 在 $\lambda(x)$比较大时，备择假设成立观察到样本点 x 的可能性比较大，因此很自然地，在 $\lambda(x)$比较大时拒绝原假设. 故取检验的拒绝域为$\{x:\lambda(x)\geqslant c\}$. 这个检验方法常用于区分样本来自这类分布，还是那类分布的检验问题.

例 3.14　设有样本 $X=(X_1,\cdots,X_n)$. 考虑检验问题，原假设和备择假设分别是 H_0：样本来自正态分布 $N(\mu,\sigma^2)$，其密度函数为

$$p_0(x;\mu,\sigma^2)=(\sqrt{2\pi})^{-1}\cdot\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

和 H_1：样本来自双参数指数分布族 $Exp(\mu,\sigma)$，其密度函数为

$$p_1(x;\mu,\sigma)=\frac{1}{\sigma}\cdot\exp\left\{-\frac{x-\mu}{\sigma}\right\}\text{或 }0\text{，在 }x-\mu\geqslant 0\text{ 或 }<0\text{ 时}$$

其中$-\infty<\mu<\infty,\sigma>0$. 原假设成立时，$\mu$ 和 σ 的 MLE 分别为

$$\hat{\mu}_0=\overline{X},\hat{\sigma}_0=\sqrt{\frac{1}{n}\cdot\sum_{i=1}^{n}(X_i-\overline{X})^2}$$

而备择假设成立时 μ 和 σ 的 MLE 分别为

$$\hat{\mu}_1=X_{(1)}=\min\{X_1,\cdots,X_n\},\hat{\sigma}_1=\overline{X}-X_{(1)}$$

所以该检验问题的似然比

$$\lambda(x)=\frac{\prod\limits_{i=1}^{n}p_0(x_i;\hat{\mu}_0,\hat{\sigma}_0^2)}{\prod\limits_{i=1}^{n}p_1(x_i;\hat{\mu}_1,\hat{\sigma}_1)}=\left(\sqrt{2\pi\cdot e^{-1}}\right)^n\cdot D^n$$

其中

$$D=\frac{\sqrt{n\cdot\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}}{\sum\limits_{i=1}^{n}(x_i-x_{(1)})}$$

由于 $\lambda(x)$ 关于 D 严格增加，所以拒绝域可取为 $\{x: D\geqslant c\}$. 由于

$$D=\frac{\sqrt{n\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2}}{\sum_{i=1}^{n}(x_i-x_{(1)})}=\frac{\sqrt{n\cdot\sum_{i=1}^{n}(u_i-\bar{u})^2}}{\sum_{i=1}^{n}(u_i-u_{(1)})}$$

其中 $u_i=\frac{x_i-\mu}{\sigma}, i=1,2,\cdots,n$，所以不论原假设为真，还是备择假设为真，$D$ 的分布皆与 μ 和 σ 无关. 原假设为真时，检验犯第一类错误的概率为

$$P\{D\geqslant c \mid \text{样本来自标准正态分布 } N(0,1)\}$$

利用随机模拟法，对不同的样本容量 n，求得原假设为真时 D 的分位数值，从而得到了检验的临界值 c. 然后由于备择假设为真时，检验犯第二类错误的概率为

$$P\{D< c \mid \text{样本来自标准指数分布 } Exp(0,1)\}$$

利用随机模拟法，可以求得检验犯第二类错误的概率.

简单原假设对简单备择假设的似然比 $\lambda(x)$ 的另一个很自然的推广为

$$\lambda(x)=\frac{\sup\limits_{\theta\in\Theta}p(x;\theta)}{\sup\limits_{\theta\in\Theta_0}p(x;\theta)}=\frac{p(x;\hat{\theta})}{p(x;\hat{\theta}_0)}$$

其中，$\hat{\theta}$ 是 $\theta\in\Theta$，即 θ 没有受到限制时的 MLE；$\hat{\theta}_0$ 是原假设 $H_0: \theta\in\Theta_0$ 成立时，即 θ 受到限制时的 MLE. 显然，$\lambda(x)\geqslant 1$. 检验的拒绝域为 $\{x: \lambda(x)\geqslant c\}$. 似然比的这种推广较前者普遍. 它常用于参数假设检验问题.

例 3.15　设样本 $X=(X_1,\cdots,X_n)$ 来自正态分布 $N(\mu,\sigma^2)$，其中 $-\infty<\mu<\infty, \sigma^2>0$. 考虑检验问题，原假设 $H_0: \mu=0$ 对备择假设 $H_1: \mu\neq 0$. 原假设 $H_0: \mu=0$ 成立时，σ^2 的 MLE 为 $\hat{\sigma}_0^2=\frac{1}{n}\cdot\sum_{i=1}^{n}X_i^2$；而备择假设 $H_1: \mu\neq 0$ 成立时，即均值 μ 和方差 σ^2 皆未知时，μ 和 σ^2 的 MLE 分别为 $\hat{\mu}=\bar{X}$ 和 $\hat{\sigma}^2=\frac{1}{n}\cdot\sum_{i=1}^{n}(X_i-\bar{X})^2$. 因而似然比

$$\lambda(x) = \frac{\prod_{i=1}^{n}(\sqrt{2\pi}\,\hat{\sigma})^{-1}\cdot\exp\left\{-\frac{(x_i-\hat{\mu})^2}{2\hat{\sigma}^2}\right\}}{\prod_{i=1}^{n}(\sqrt{2\pi}\,\hat{\sigma}_0)^{-1}\cdot\exp\left\{-\frac{x_i^2}{2\hat{\sigma}_0^2}\right\}}$$

$$= \left[\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2 + n\cdot\overline{x}^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right]^{\frac{n}{2}} = \left[1+\frac{t^2}{n-1}\right]^{\frac{n}{2}}$$

其中

$$t = \sqrt{n(n-1)}\frac{\overline{x}}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}}$$

显然，$\lambda(x)$是$|t|$的严增函数. 故拒绝域可取为$\{x: |t|\geqslant c\}$. 原假设$H_0: \mu=0$成立时，$t\sim t(n-1)$. 故取$c=t_{1-\alpha/2}(n-1)$. 这就是常用的t检验法. 不仅t检验如此，前面几节给出的不少最优检验都是似然比检验. 由于似然比检验的适用面广，且由它构造出来的检验常具有某种优良性，所以似然比检验是假设检验中的一个重要方法.

一般来说，难以求得似然比统计量的精确分布. 因而有必要寻求，在样本容量趋于无穷时，似然比统计量的极限分布. 1938 年，Wilks 证明了关于似然比统计量的极限分布的重要定理. 下面分简单原假设和复合原假设两种情况进行讨论.

§3.6.2　简单原假设的检验问题

定理 3.18　设样本$X=(X_1,\cdots,X_n)$来自密度函数族$\{p(x;\theta): \theta\in\Theta\}$，其中未知参数$\theta=(\theta_1,\cdots,\theta_k)$，参数空间$\Theta$是$k$维欧氏空间$\mathbf{R}^k$中的一个含有内点的集合. 假设参数真值$\theta_0$是$\Theta$的一个内点.

假设该密度函数族满足下列四个条件：

(1) $\int_{\mathscr{X}}\frac{\partial p(x;\theta)}{\partial\theta_i}\mathrm{d}\mu(x) = \int_{\mathscr{X}}\frac{\partial^2 p(x;\theta)}{\partial\theta_i\partial\theta_j}\mathrm{d}\mu(x) = 0,\quad i,j=1,\cdots,k$

(2) $\boldsymbol{I}(\theta)=(I_{ij}(\theta))_{k\times k}>0,\quad \forall\,\theta\in\Theta$，其中

$$I_{ij}(\theta)=\int_{\mathcal{X}}\left(\frac{\partial\ln p(x;\theta)}{\partial\theta_i}\right)\cdot\left(\frac{\partial\ln p(x;\theta)}{\partial\theta_j}\right)\cdot p(x;\theta)\mathrm{d}\mu(x)$$

(3)存在 $M(x)$，使得 $\int_{\mathcal{X}}M(x)\cdot p(x;\theta)\mathrm{d}\mu(x)<K,\forall\ \theta\in\Theta$，其中 K 与 θ 无关，且在含有参数真值 θ_0 的一个邻域内，

$$\left|\frac{\partial^3\ln p(x;\theta)}{\partial\theta_i\partial\theta_j\partial\theta_l}\right|\leqslant M(x),\quad i,j,l=1,\cdots,k$$

(4)不同的 θ 值，对应着不同的概率分布.

除上述四个条件外，再假设 θ 的 MLE$\hat{\theta}$ 是似然方程 $\frac{\partial l(\theta)}{\partial\theta}=0$ 的解，其中

$$l(\theta)=\ln\prod_{i=1}^{n}p(x_i;\theta)=\sum_{i=1}^{n}\ln p(x_i;\theta)$$

并且在 $n\to\infty$ 时，$\hat{\theta}$ 依概率收敛于参数真值 θ_0.

考虑如下双边检验问题：

简单原假设 $H_0:\theta=\theta_0$ 对 备择假设 $H_1:\theta\neq\theta_0$

该检验问题的似然比统计量为

$$\lambda(X)=\frac{\prod_{i=1}^{n}p(X_i;\hat{\theta})}{\prod_{i=1}^{n}p(X_i;\theta_0)}$$

那么在原假设 H_0 成立时，$2\ln\lambda(X)$ 随 n 增大而依分布收敛于 $\chi^2(k)$.

证明：下面的讨论都是在原假设 H_0 成立的条件下进行的. 由于 $\hat{\theta}$ 是似然方程的相合解，则据 MLE 的渐近正态性定理可知，

$$D(\theta_0)\xrightarrow{L}N(0,\boldsymbol{I}^{-1}(\theta_0)),\quad n\to\infty \tag{3.61}$$

其中 $D(\theta_0)=\sqrt{n}(\hat{\theta}-\theta_0)$. 不难证明(习题 3.17)

$$I_{ij}(\theta)=\mathrm{E}_\theta\left[\left(\frac{\partial\ln p(X;\theta)}{\partial\theta_i}\right)\cdot\left(\frac{\partial\ln p(X;\theta)}{\partial\theta_j}\right)\right]=\mathrm{E}_\theta\left[\frac{\partial^2\ln p(X;\theta)}{\partial\theta_i\partial\theta_j}\right]$$

将 $l(\theta_0)=\sum_{i=1}^{n}\ln p(x_i;\theta_0)$ 在 $\hat{\theta}$ 处泰勒展开. 由于 $\hat{\theta}$ 是似然方程的解，则有

$$l(\theta_0)=l(\hat{\theta})+(\theta_0-\hat{\theta})'\cdot l'(\hat{\theta})+\frac{1}{2}\cdot(\theta_0-\hat{\theta})'\cdot l''(\theta^*)\cdot(\theta_0-\hat{\theta})$$

$$= l(\hat{\theta}) - \frac{1}{2} \cdot D(\theta_0)' \cdot C \cdot D(\theta_0)$$

其中

$$C = (c_{ij}(\theta^*))_{k\times k}, \quad c_{ij}(\theta) = -\frac{1}{n} \cdot \sum_{r=1}^{n} \frac{\partial^2 \ln p(x_r;\theta)}{\partial\theta_i \partial\theta_j}$$

这里 $\theta^* = \hat{\theta} + \eta(\hat{\theta} - \theta_0)$，$0 \leqslant \eta \leqslant 1$. 则有

$$2\ln\lambda(x) = 2\left(\sum_{i=1}^{n} \ln p(x_i;\hat{\theta}) - \sum_{i=1}^{n} \ln p(x_i;\theta_0) \right)$$

$$= 2(l(\hat{\theta}) - l(\theta_0)) = D(\theta_0)' \cdot C \cdot D(\theta_0) \qquad (3.62)$$

将 $c_{ij}(\theta^*)$在 θ_0 处泰勒展开，并应用定理关于三阶偏导数的假设条件，则有

$$c_{ij}(\theta^*) = c_{ij}(\theta_0) + (\theta^* - \theta_0)' \cdot \left[\frac{\partial c_{ij}(\theta)}{\partial\theta}\right]_{\theta=\theta_0 - v\cdot(\theta_0 - \theta^*)}$$

$$\leqslant c_{ij}(\theta_0) + \left(\sum_{l=1}^{k} |\theta_l^* - \theta_{0l}| \right) \cdot \frac{\sum_{r=1}^{n} M(x_r)}{n}$$

其中 $0 \leqslant v \leqslant 1$，$\theta^* = (\theta_1^*, \cdots, \theta_k^*)$，$\theta_0 = (\theta_{01}, \cdots, \theta_{0k})$. 由于 $\hat{\theta}$ 随 n 增大而依概率收敛于参数真值 θ_0. 从而 $\theta^* \xrightarrow{p} \theta_0$. 故对每一个 l，都有 $\theta_l^* \xrightarrow{p} \theta_{0l}$. 又由大数定律知，$\frac{1}{n} \cdot \sum_{r=1}^{n} M(X_r) \xrightarrow{p} E_{\theta_0}[M(X)] < K$，所以 $c_{ij}(\theta^*) \xrightarrow{p} c_{ij}(\theta_0)$. 由大数定律知

$$c_{ij}(\theta_0) \xrightarrow{p} -E_{\theta_0}\left[\frac{\partial^2 \ln p(X;\theta)}{\partial\theta_i \partial\theta_j}\right] = I_{ij}(\theta_0)$$

所以 $c_{ij}(\theta^*) \xrightarrow{p} I_{ij}(\theta_0)$. 从而

$$C \xrightarrow{p} I(\theta_0) \qquad (3.63)$$

则由(3.62)式知，在 $n \to \infty$ 时，$2\ln\lambda(X)$ 与 $D(\theta_0)' \cdot I(\theta_0) \cdot D(\theta_0)$有相同的极限分布. 由(3.61)式知，在 $n \to \infty$ 时，$I^{\frac{1}{2}}(\theta_0) \cdot D(\theta_0) \xrightarrow{L} N(0,1)$. 故 $D(\theta_0)' \cdot I(\theta_0) \cdot D(\theta_0)$，从而 $2\ln\lambda(X)$随 n 增大而依分布收敛于 $\chi^2(k)$. 定理证毕.

由(3.62)和(3.63)两式知,

$$2\ln\lambda(X) = 2\left(\sum_{i=1}^{n}\ln p(X_i;\hat{\theta}) - \sum_{i=1}^{n}\ln p(X_i;\theta_0)\right)$$
$$= D(\theta_0)' \cdot C \cdot D(\theta_0)$$
$$= D(\theta_0)' \cdot I(\theta_0) \cdot D(\theta_0) + \xi_n \tag{3.64}$$

其中 $\xi_n \xrightarrow{p} 0$. 这说明,$2\ln\lambda(X)$ 与 $D(\theta_0)' \cdot I(\theta_0) \cdot D(\theta_0)$ 仅相差一个依概率收敛于 0 的量. 在将定理 3.18 推广到复合原假设的情况时,(3.64)式起着重要的作用.

例 3.16 分类数据的检验. 根据某项指标,总体被分成 r 类:A_1,$\cdots$,A_r. 由经验或某个理论提出了如下的检验问题:

原假设 H_0:随机抽取一个个体 x 进行观测,$P(x\in A_i)$已知,等于 $p_{0i}(i=1,\cdots,r)$,其中

$$\sum_{i=1}^{r} p_{0i} = 1$$

备择假设 H_1:$P(x\in A_i)=p_{0i}(i=1,\cdots,r)$不全成立.

随机抽取 n 个个体进行观测. 设有 n_i 个个体属于类 A_i. 则有 $\sum_{i=1}^{r} n_i = n$. 样本$\{n_i: i=1,\cdots,r\}$服从多项分布,其概率密度为

$$p(n_1,\cdots,n_r) = \frac{n!}{n_1!\cdots n_r!} \cdot \prod_{i=1}^{r} p_i^{n_i}$$

这里 p_i 是类 A_i 所占的比例. 由于 $\sum_{i=1}^{n} p_i=1$,所以参数空间

$$\Theta = \{(p_1,\cdots,p_r): p_i \geqslant 0, i=1,\cdots,r; \sum_{i=1}^{r} p_i = 1\}$$

中独立参数只有 $r-1$ 个. 故 Θ 是 $r-1$ 维欧氏空间的一个含有内点的集合. p_i 的 MLE 为 $\hat{p}_i=n_i/n$. 该检验问题的似然比统计量为

$$\Lambda = \frac{\prod_{i=1}^{r}\left(\frac{n_i}{n}\right)^{n_i}}{\prod_{i=1}^{r}(p_{0i})^{n_i}}$$

据定理 3.18 知,在原假设 H_0 为真时,$2\ln\Lambda = 2\sum_{i=1}^{r} n_i \cdot \ln\frac{n_i}{np_{0i}}$

$\xrightarrow{L}\chi^2(r-1)$. 故在 $2\sum_{i=1}^{r} n_i \cdot \ln \frac{n_i}{np_{0i}} \geqslant \chi^2_{1-\alpha}(r-1)$时，拒绝原假设.

在原假设 H_0 为真时，可以证明(见习题 3.19)

$$
\begin{aligned}
2\ln\Lambda &= 2\sum_{i=1}^{r} n_i \cdot \ln\left[1+\frac{n_i-np_{0i}}{np_{0i}}\right] \\
&= 2\sum_{i=1}^{r} n_i \cdot \left[\frac{n_i-np_{0i}}{np_{0i}}-\frac{1}{2}\cdot\left(\frac{n_i-np_{0i}}{np_{0i}}\right)^2\right] \\
&\quad + \text{依概率收敛于 0 的量} \\
&= 2\sum_{i=1}^{r}[np_{0i}+(n_i-np_{0i})]\cdot\left[\frac{n_i-np_{0i}}{np_{0i}}-\frac{1}{2}\cdot\left(\frac{n_i-np_{0i}}{np_{0i}}\right)^2\right] \\
&\quad + \text{依概率收敛于 0 的量} \\
&= \sum_{i=1}^{r}\frac{(n_i-np_{0i})^2}{np_{0i}} + \text{依概率收敛于 0 的量}
\end{aligned}
$$

右边就是分类数据检验问题的 Pearsonχ^2 拟合优度检验统计量. 所以 χ^2 拟合优度检验可以近似看作分类数据的似然比检验.

§3.6.3　复合原假设的检验问题

设样本 $X=(X_1,\cdots,X_n)$来自密度函数族$\{p(x;\theta):\theta\in\Theta\}$，其中未知参数 $\theta=(\theta_1,\cdots,\theta_k)$，参数空间 Θ 是 k 维欧氏空间 $\mathbf{R}^k$ 中的一个含有内点的集合. 考虑复合原假设检验问题：原假设 $H_0:\theta\in\Theta_0$ 对备择假设 $H_1:\theta\notin\Theta_0$，其中 Θ_0 是 Θ 的一个非单点子集. 一般来说，复合原假设大致有以下几种情况：

情况 1. 原假设 $H_0:\theta_{r+1}=\cdots=\theta_k=0$.

这时，$\Theta_0=\{\theta:\theta=(\theta_1,\cdots,\theta_r,0,\cdots,0)\in\Theta\}$. 例如，总体为 $N(\mu,\sigma^2)$，其中 $-\infty<\mu<\infty$，$\sigma^2>0$. 这时，$\theta=(\mu,\sigma^2)\in\mathbf{R}\times\mathbf{R}^+$. 复合原假设 $H_0:\mu=0$ 就属于情况 1. 情况 1 最为简单. 下述情况 2 是情况 1 的推广.

情况 2. 原假设 $H_0:R_i(\theta_1,\cdots,\theta_k)=0,i=1,\cdots,s$，其中 $R_i(i=1,\cdots,s)$具有一阶连续偏导数. 这时，参数 θ 受到 s 个条件的限制，$\Theta_0=\{\theta:\theta\in\Theta;R_i(\theta)=0,i=1,\cdots,s\}$. 例如，相互独立的两个总体 $N(\mu_1,\sigma^2)$和

$N(\mu_2,\sigma^2)$，其中 $-\infty<\mu_1,\mu_2<\infty,\sigma^2>0$. 这时 $\theta=(\mu_1,\mu_2,\sigma^2)\in\mathbf{R}\times\mathbf{R}\times\mathbf{R}^+$. 复合原假设 $H_0:\mu_1=\mu_2$ 就属于情况 2. 事实上，可适当地重新定义参数 $\boldsymbol{\omega}=(\omega_1,\cdots,\omega_k)$，使得原假设变换为 $H_0:\omega_{k-r+1}=\cdots=\omega_k=0$. 则情况 2 就变成情况 1. 在上例中，若用新参数 $(\mu_1,\mu_1-\mu_2,\sigma^2)$ 取代原来的参数 (μ_1,μ_2,σ^2)，则就变成情况 1.

情况 3. 原假设 $H_0:\theta_i=g_i(\varphi_1,\cdots,\varphi_r),i=1,\cdots,k$，其中 $g_i(i=1,\cdots,k)$ 具有一阶连续偏导数. 这时，$\Theta_0=\{\theta:\theta\in\Theta;\theta_i=g_i(\varphi),i=1,\cdots,k,\varphi=(\varphi_1,\cdots,\varphi_r)\in A\}$，$A$ 是 r 维欧氏空间 $\mathbf{R}^r$ 中的一个含有内点的集合. 例如，设有 $\boldsymbol{X},\boldsymbol{Y}$ 二个离散型随机变量. $\boldsymbol{X}$ 表示人的性别，"$\boldsymbol{X}=1$"表示男性；"$\boldsymbol{X}=2$"表示女性；$\boldsymbol{Y}$ 表示是否色盲，"$Y=1$"表示正常；"$\boldsymbol{Y}=2$"表示色盲. 令 $p_{ij}=P(X=i,Y=j),i,j=1,2$. 这时 $\boldsymbol{\theta}=(p_{11},p_{12},p_{21})$，其中，$0\leqslant p_{11},p_{12},p_{21}\leqslant1$，且 $p_{11}+p_{12}+p_{21}\leqslant1$. 所以参数空间 $\boldsymbol{\Theta}$ 为 $\mathbf{R}^3$ 中的一个单纯形.

有 1000 人按性别与色盲分类如下

	正常	色盲
男	442	38
女	514	6

按遗传学模型，数据应有下列相对应的概率，

	正常	色盲
男	$\frac{p}{2}$	$\frac{1-p}{2}$
女	$\frac{p^2}{2}+p\cdot(1-p)$	$\frac{(1-p)^2}{2}$

其中 $0\leqslant p\leqslant1$. 问数据与模型是否相符？这里所要检验的原假设为

$$H_0: p_{11}=\frac{p}{2},\quad p_{12}=\frac{1-p}{2},\quad p_{21}=\frac{p^2}{2}+p\cdot(1-p)$$

这属于情况 3.

取 $\varphi=(\theta_1,\cdots,\theta_r)$，则情况 1 变换成情况 3. 从而情况 2 也能变换成情况 3. 事实上，情况 3 也可以变换成情况 2. 如果能将 $\varphi_1,\cdots,\varphi_s$ 表示成 $\theta_1,\cdots,\theta_s$ 的函数，然后把这些函数代入 $\theta_j=g_j(\varphi_1,\cdots,\varphi_s)$，$j=s+1,\cdots,k$，由此可得到参数 θ 受到的 $k-s$ 个条件的限制，则情况 3 变换成情况 2. 在上例中，由于 $p_{11}=p/2$，所以 $p=2\cdot p_{11}$. 因此参数 $\theta=(p_{11},p_{12},p_{21})$受到两个条件的限制：$p_{12}=(1-2\cdot p_{11})/2$，以及 $p_{21}=2\cdot p_{11}^2+2\cdot p_{11}\cdot(1-2\cdot p_{11})$.

下面我们只讨论情况 3 的复合原假设的检验问题.

定理 3.19 设样本 $X=(X_1,\cdots,X_n)$来自密度函数族$\{p(x;\theta):\theta\in\Theta\}$，其中未知参数 $\theta=(\theta_1,\cdots,\theta_k)$，参数空间 Θ 是 k 维欧氏空间 $\mathbf{R}^k$ 中的一个含有内点的集合. 参数真值 θ_0 是 Θ 的一个内点. 假设该密度函数族满足定理 3.18 的四个条件.

考虑原假设 $H_0:\theta\in\Theta_0$ 对备择假设 $H_1:\theta\notin\Theta_0$，其中 Θ_0 是 Θ 的一个子集，且满足下述条件：存在 $r(r<k)$维欧氏空间 $\mathbf{R}^r$ 中的一个含有内点的集合 A，以及定义在 A 上的 k 个三阶可导的函数 $g=(g_1,\cdots,g_k)$，使得 A 与 Θ_0 一一对应，

$$\Theta_0=\{\theta:\theta=g(\varphi),\varphi\in A\}$$

在参数真值 $\theta_0\in\Theta_0$ 时，假设 θ_0 对应于 A 中的点 $\varphi_0:\theta_0=g(\varphi_0)$，$\varphi_0$ 是 A 的内点.

在 $\theta=g(\varphi)$时，记 $p(x:\theta)=p(x;g(\varphi))$为 $\tilde{p}(x;\varphi)$. 则分布族$\{p(x;\theta):\theta\in\Theta_0\}$可等价地写成$\{\tilde{p}(x;\varphi):\varphi\in A\}$的形式. 假设密度函数族$\{\tilde{p}(x;\varphi):\varphi\in A\}$满足定理 3.18 的四个条件.

在 $\theta\in\Theta$ 时，假设 θ 的 MLE$\hat{\theta}$ 是似然方程$\frac{\partial l(\theta)}{\partial\theta}=0$ 的解，其中

$$l(\theta)=\ln\prod_{i=1}^n p(x_i;\theta)=\sum_{i=1}^n\ln p(x_i;\theta)$$

并且在 $n\to\infty$时，$\hat{\theta}$ 依概率收敛于参数真值 θ_0. 在原假设为真，即参数

真值 $\theta_0\in\Theta_0$ 时，θ 的 MLE$\hat{\theta}_0=g(\hat{\varphi}_0)$，假设 $\hat{\varphi}_0$ 是似然方程 $\frac{\partial\tilde{l}(\varphi)}{\partial(\varphi)}=0$ 的解，其中

$$\tilde{l}(\varphi)=\ln\prod_{i=1}^{n}\tilde{p}(x_i;\varphi)=\sum_{i=1}^{n}\ln\tilde{p}(x_i;\varphi)$$

并且在 $n\to\infty$ 时，$\hat{\varphi}_0$ 依概率收敛于 φ_0.

定义似然比统计量

$$\lambda(X)=\frac{\prod_{i=1}^{n}p(X_i;\hat{\theta})}{\prod_{i=1}^{n}p(X_i;\hat{\theta}_0)}$$

则在原假设 H_0 成立时，$2\ln\lambda(X)$ 随 n 增大而依分布收敛于 $\chi^2(k-r)$.

证明：下面的讨论都是在原假设 H_0 成立的条件下进行的，并且沿袭定理 3.18 的符号.

令

$$V(\theta)=\frac{1}{\sqrt{n}}\cdot\left(\sum_{i=1}^{n}\frac{\partial p(x_i;\theta)}{\partial\theta_1},\cdots,\sum_{i=1}^{n}\frac{\partial p(x_i;\theta)}{\partial\theta_k}\right)'$$

则由中心极限定理知，在 $n\to\infty$ 时，

$$V(\theta_0)\xrightarrow{L}N(0,I(\theta_0))\tag{3.65}$$

从而由(3.61)式知，$I^{-1}(\theta_0)\cdot V(\theta_0)$ 与 $D(\theta_0)$ 有相同的极限分布. 不仅如此，由 MLE 渐近正态性定理的证明过程知，$I^{-1}(\theta_0)\cdot V(\theta_0)$ 与 $D(\theta_0)$ 仅相差一个依概率收敛于 0 的量.

因为密度函数族 $\{p(x;\theta):\theta\in\Theta\}$ 满足定理 3.18 的四个条件，则由(3.64)式知，

$$2\left(\sum_{i=1}^{n}\ln p(X_i;\hat{\theta})-\sum_{i=1}^{n}\ln p(X_i;\theta_0)\right)\text{ 与 }D(\theta_0)'\cdot I(\theta_0)\cdot D(\theta_0)$$

从而

$$2\left(\sum_{i=1}^{n}\ln p(X_i;\hat{\theta})-\sum_{i=1}^{n}\ln p(X_i;\theta_0)\right)\text{ 与}[V(\theta_0)]'\cdot[I(\theta_0)]^{-1}\cdot V(\theta_0)$$

仅相差一个依概率收敛于 0 的量，它们有相同的极限分布.

同理，因为密度函数族 $\{\tilde{p}(x;\varphi):\varphi\in A\}$ 也满足定理 3.18 的四个

条件，所以

$$2\left\{\sum_{i=1}^{n}\ln\tilde{p}(X_i;\hat{\varphi})-\sum_{i=1}^{n}\ln\tilde{p}(X_i;\boldsymbol{\varphi}_0)\right\}\text{与}[\boldsymbol{U}(\boldsymbol{\varphi}_0)]'\cdot[\boldsymbol{J}(\boldsymbol{\varphi}_0)]^{-1}\cdot\boldsymbol{U}(\boldsymbol{\varphi}_0)$$

仅相差一个依概率收敛于 0 的量，它们有相同的极限分布，其中

$$\boldsymbol{U}(\boldsymbol{\varphi})=\frac{1}{\sqrt{n}}\cdot\left(\sum_{i=1}^{n}\frac{\partial\tilde{p}(x_i;\boldsymbol{\varphi})}{\partial\varphi_1},\cdots,\sum_{i=1}^{n}\frac{\partial\tilde{p}(x_i;\boldsymbol{\varphi})}{\partial\varphi_r}\right)$$

$$\boldsymbol{J}(\boldsymbol{\varphi})=(J_{ij}(\boldsymbol{\varphi}))_{r\times r}$$

其中

$$J_{ij}(\boldsymbol{\varphi})=\int_{\mathscr{X}}\left(\frac{\partial\ln\tilde{p}(x;\boldsymbol{\varphi})}{\partial\varphi_i}\right)\cdot\left(\frac{\partial\ln\tilde{p}(x;\boldsymbol{\varphi})}{\partial\varphi_j}\right)\cdot\tilde{p}(x;\boldsymbol{\varphi})\mathrm{d}\mu(x)$$

令 $\boldsymbol{B}(\boldsymbol{\varphi})=(B_{ij}(\boldsymbol{\varphi}))_{r\times r}$，其中 $B_{ij}(\boldsymbol{\varphi})=\dfrac{\partial g_j(\boldsymbol{\varphi})}{\partial\varphi_i}$. 注意到 $\theta=\boldsymbol{g}(\boldsymbol{\varphi})$，$\boldsymbol{\theta}_0=\boldsymbol{g}(\boldsymbol{\varphi}_0)$. 从而有

$$\boldsymbol{U}(\boldsymbol{\varphi}_0)=\boldsymbol{B}(\boldsymbol{\varphi}_0)\cdot\boldsymbol{V}(\theta_0)$$

$$\boldsymbol{J}(\boldsymbol{\varphi}_0)=\boldsymbol{B}(\boldsymbol{\varphi}_0)\cdot\boldsymbol{I}(\theta_0)\cdot[\boldsymbol{B}(\boldsymbol{\varphi}_0)]' \tag{3.66}$$

所以

$$2\left\{\sum_{i=1}^{n}\ln\tilde{p}(X_i;\hat{\varphi})-\sum_{i=1}^{n}\ln\tilde{p}(X_i;\varphi_0)\right\}$$

与

$$[\boldsymbol{V}(\theta_0)]'\cdot[\boldsymbol{B}(\boldsymbol{\varphi}_0)]'\cdot[\boldsymbol{J}(\boldsymbol{\varphi}_0)]^{-1}\cdot\boldsymbol{B}(\boldsymbol{\varphi}_0)\cdot\boldsymbol{V}(\theta_0)$$

仅相差一个依概率收敛于 0 的量，它们有相同的极限分布.

注意到 $p(x;\theta_0)=\tilde{p}(x;\boldsymbol{\varphi}_0)$，$p(x;\theta_0)=\tilde{p}(x;\boldsymbol{\varphi}_0)$，从而有

$$\begin{aligned}2\ln\lambda(X)&=2\left\{\sum_{i=1}^{n}\ln p(X_i;\hat{\theta})-\sum_{i=1}^{n}\ln p(X_i;\hat{\theta}_0)\right\}\\&=2\left\{\sum_{i=1}^{n}\ln p(X_i;\hat{\theta})-\sum_{i=1}^{n}\ln p(X_i;\theta_0)\right\}-\\&\quad 2\left\{\sum_{i=1}^{n}\ln\tilde{p}(X_i;\hat{\boldsymbol{\varphi}}_0)-\sum_{i=1}^{n}\ln\tilde{p}(X_i;\boldsymbol{\varphi}_0)\right\}\end{aligned}$$

则 $2\ln\lambda(X)$ 与

$$[\boldsymbol{V}(\theta_0)]'\cdot\{[\boldsymbol{I}(\theta_0)]^{-1}-[\boldsymbol{B}(\boldsymbol{\varphi}_0)]'\cdot[\boldsymbol{J}(\boldsymbol{\varphi}_0)]^{-1}\cdot\boldsymbol{B}(\boldsymbol{\varphi}_0)\}\cdot\boldsymbol{V}(\boldsymbol{\theta}_0)$$

仅相差一个依概率收敛于 0 的量，它们有相同的极限分布. 令 $\boldsymbol{W}(\boldsymbol{\theta}_0)=$

$[\boldsymbol{I}(\theta_0)]^{-\frac{1}{2}}\cdot \boldsymbol{V}(\theta_0)$，则由(3.65)式知，$\boldsymbol{W}(\theta_0)\xrightarrow{L}N(0,\boldsymbol{I})$，且 $2\ln\lambda(\boldsymbol{X})$与

$$[\boldsymbol{W}(\theta_0)]'\cdot[\boldsymbol{I}-\boldsymbol{G}(\theta_0)]\cdot \boldsymbol{W}(\theta_0)$$

有相同的极限分布，其中

$$\boldsymbol{G}(\theta_0)=[\boldsymbol{I}(\theta_0)]^{\frac{1}{2}}\cdot[\boldsymbol{B}(\varphi_0)]'\cdot[\boldsymbol{J}(\varphi_0)]^{-1}\cdot \boldsymbol{B}(\varphi_0)\cdot[\boldsymbol{I}(\theta_0)]^{\frac{1}{2}}$$

由(3.66)式知，$\boldsymbol{I}-\boldsymbol{G}(\theta_0)$是幂等阵，且$\boldsymbol{I}-\boldsymbol{G}(\theta_0)$的秩为

$$\begin{aligned}k-\operatorname{tr}\{\boldsymbol{G}(\theta_0)\}&=k-\operatorname{tr}\{[\boldsymbol{J}(\varphi_0)]^{-1}\cdot \boldsymbol{B}(\varphi_0)\cdot[\boldsymbol{I}(\theta_0)]\cdot[\boldsymbol{B}(\varphi_0)]'\}\\&=k-r\end{aligned}$$

所以 $2\ln\lambda(\boldsymbol{X})$的极限分布为 $\chi^2(k-r)$. 定理证毕.

§3.6.4 二维列联表的独立性检验

设有 X,Y 二个离散型随机变量，分别取 r 个和 c 个值，与此相联系的有一个二维列联表 $r\times c$. 作 n 次观测，在(i,j)格的观测频数为 n_{ij}，$i=1,2,\cdots,r$；$j=1,2,\cdots,c$. 观测值落入(i,j)格的概率为 p_{ij}. 观测频数$\{n_{ij},i=1,2,\cdots,r;j=1,2,\cdots,c\}$服从多项分布，其概率密度为

$$\frac{n!}{\prod_{i=1}^{r}\prod_{j=1}^{c}n_{ij}!}\cdot\prod_{i=1}^{r}\prod_{j=1}^{c}p_{ij}^{n_{ij}}$$

由于 $\sum_{i=1}^{r}\sum_{j=1}^{c}p_{ij}=1$，所以参数空间

$$\Theta=\left\{p_{ij}:i=1,\cdots,r;j=1,\cdots,c;p_{ij}\geqslant 0,\sum_{i=1}^{r}\sum_{j=1}^{c}p_{ij}=1\right\}$$

中独立参数只有 $r\cdot c-1$ 个. 故 Θ 是 $r\cdot c-1$ 维欧氏空间的一个含有内点的集合. p_{ij}的 MLE 为 $\hat{p}_{ij}=n_{ij}/n$.

考虑独立性检验，原假设 H_0：$p_{ij}=p_{i\cdot}\cdot p_{\cdot j}$，$i=1,2,\cdots,r$；$j=1,2,\cdots,c$，其中 $p_{i\cdot}$和 $p_{\cdot j}$分别是 X 和 Y 边缘分布. 原假设成立时，$p_{ij}=p_{i\cdot}\cdot p_{\cdot j}$，所以其参数空间 Θ_0 与下述集合 A 一一对应：

$$A=\left\{\begin{aligned}&(p_{i\cdot},p_{\cdot j}):i=1,\cdots,r;j=1,\cdots,c,\\&p_{i\cdot}\geqslant 0,p_{\cdot j}\geqslant 0,\sum_{i=1}^{r}p_{i\cdot}=\sum_{j=1}^{c}p_{\cdot j}=1\end{aligned}\right\}$$

A 中独立参数只有 $r+c-2$ 个. 故 A 是 $r+c-2$ 维欧氏空间的一个含

有内点的集合. 这时, p_{ij} 的 MLE 为 $\hat{p}_{i.}\cdot\hat{p}_{.j}=(n_{i.}/n)\cdot(n_{.j}/n)$, 其中 $n_{i.}=\sum_{j=1}^{c}n_{ij}, n_{.j}=\sum_{i=1}^{r}n_{ij}$.

该检验问题的似然比统计量为

$$\Lambda=\frac{\prod_{i=1}^{r}\prod_{j=1}^{c}\left(\frac{n_{ij}}{n}\right)^{n_{ij}}}{\prod_{i=1}^{r}\prod_{j=1}^{c}\left(\frac{n_{i.}}{n}\cdot\frac{n_{.j}}{n}\right)^{n_{ij}}}$$

由于 $(r\cdot c-1)-(r+c-2)=(r-1)(c-1)$, 故在原假设成立时, $2\ln\Lambda$ 的极限分布为 $\chi^2((r-1)(c-1))$. 在 $2\ln\Lambda\geqslant\chi^2_{1-\alpha}((r-1)(c-1))$ 时, 也就是在

$$2\sum_{i=1}^{r}\sum_{j=1}^{c}n_{ij}\cdot\ln\frac{n\cdot n_{ij}}{n_{i.}\cdot n_{.j}}\geqslant\chi^2_{1-\alpha}((r-1)(c-1))$$

的时候, 拒绝原假设.

与例 3.15 相类似地, 在原假设 H_0 为真时, 可以推出(见习题 3.19)

$$2\ln\Lambda=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-n\hat{p}_{i.}\hat{p}_{j.})^2}{n\hat{p}_{i.}\hat{p}_{j.}}+\text{依概率收敛于 0 的量}$$

右边就是二维列联表的独立性检验问题的 χ^2 拟合优度检验统计量.

§3.6.5　三维列联表的条件独立性检验

首先看下面的例子. 表 3.2 是 1976 年～1977 年美国佛罗里达州的凶杀案件中, 326 个被告的肤色和死刑判决情况的分类表[6].

表 3.2　被告的肤色和死刑判决情况的分类表

被告	死刑 是	死刑 否	合计
白人	19	141	160
黑人	17	149	166
合计	36	290	326

二维列联表的独立性检验似然比统计量 $2\ln\Lambda=0.2214$, 它比 $\chi^2(1)$

分布的中位数 0.4550 还要小，不能拒绝被告的肤色和死刑判决相互独立的原假设，认为白人被判死刑的概率等于黑人被判死刑的概率.事实上，表 3.2 的数据显示出，白人被判死刑的比例较高，等于 11.9%，而黑人被判死刑的比例却较低，等于 10.2%.这显然与当时美国的现实是不相符的.后来人们将被害人的肤色也考虑进去，构造了三维列联表，得到的结论就不一样了.三维列联表如下所示.

表 3.3 被告与被害人的肤色以及死刑判决情况的分类表

被告	被害人	死刑 是	死刑 否	死刑判决的比例
白人	白人	19	132	0.126
	黑人	0	9	0.000
黑人	白人	11	52	0.175
	黑人	6	97	0.058

如果被害人是白人，那么白人被判死刑的比例与黑人被判死刑的比例分别为 12.6% 与 17.5%；如果被害人是黑人，那么白人被判死刑的比例与黑人被判死刑的比例分别为 0 与 5.8%.这说明，无论被害人是白人还是黑人，黑人被判死刑的比例都是比较高的.尤其是在被害人是黑人时，白人没有被判死刑的.这说明在美国有种族歧视，死刑判决与肤色密切有关.此例告诉我们，在将三维列联表合并成二维列联表进行分析时，由于删去了一个因子，有可能产生问题.人们有必要研究三维列联表，或更高维列联表的统计分析.

设有 X,Y,Z 三个离散型随机变量，与此相联系的有一个三维列联表 $r\times c\times t$.作 n 次观测，在(i,j,k)格的观测频数为 $n_{ijk},i=1,\cdots,r;j=1,\cdots,c,k=1,\cdots,t$.观测值落入$(i,j,k)$格的概率为 p_{ijk}.$\{n_{ijk},i=1,\cdots,r;j=1,\cdots,c,l=1,\cdots,t\}$服从多项分布，其概率密度为

$$\frac{n!}{\prod_{i=1}^{r}\prod_{j=1}^{c}\prod_{k=1}^{t}n_{ijk}!}\cdot\prod_{i=1}^{r}\prod_{j=1}^{c}\prod_{k}^{t}p_{ijk}^{n_{ijk}}$$

由于 $\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{t}p_{ijk}=1$，所以参数空间

$$\Theta=\left\{p_{ijk}: i=1,\cdots,r;j=1,\cdots,c;k=1,\cdots,t;\right.$$
$$\left.p_{ijk}\geqslant 0,\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{t}p_{ijk}=1\right\}$$

中独立参数只有 $r\cdot c\cdot t-1$ 个. 故 Θ 是 $r\cdot c\cdot t-1$ 维欧氏空间的一个含有内点的集合. p_{ijk}的 MLE 为 $\hat{p}_{ijk}=n_{ijk}/n$.

这里有三个检验问题，原假设分别是三个随机变量相互独立，一个随机变量和另两个随机变量相互独立，以及给定一个随机变量后另两个随机变量相互条件独立. 前两个问题类似于二维列联表的独立性检验问题，故从略. 最后一个问题是有关条件独立性的检验，下面来讨论这个检验问题. 这里仅考虑给定 $\boldsymbol{Z}$ 之后，$\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 相互条件独立的检验问题.

若条件独立性成立，则有

$$P(\boldsymbol{X}=i,\boldsymbol{Y}=j|\boldsymbol{Z}=k)=P(\boldsymbol{X}=i|\boldsymbol{Z}=k)\cdot P(\boldsymbol{Y}=j|\boldsymbol{Z}=k)$$

从而有

$$p_{ijk}=\frac{p_{i.k}\cdot p_{.jk}}{p_{..k}} \tag{3.67}$$

其中 $p_{i.k}$，$p_{.jk}$和 $p_{..k}$分别是$(\boldsymbol{X},\boldsymbol{Z})$，$(\boldsymbol{Y},\boldsymbol{Z})$和 $\boldsymbol{Z}$ 边缘分布. 由此(习题3.21)推得，条件独立性成立时，p_{ijk}的 MLE 为

$$\frac{\hat{p}_{i.k}\cdot\hat{p}_{.jk}}{\hat{p}_{..k}}=\frac{n_{i.k}\cdot n_{.jk}}{n\cdot n_{..k}}$$

条件独立性的原假设成立时，由(3.67)式知，其参数空间 Θ_0 与下述集合 A 一一对应：

$$A=\left\{(p_{i.k},p_{.jk},p_{..k}): i=1,\cdots,r;j=1,\cdots,c;k=1,\cdots,t;\right.$$
$$\left.p_{i.k}\geqslant 0,p_{.jk}\geqslant 0,\sum_{i=1}^{r}p_{i.k}=\sum_{j=1}^{c}p_{.jk}=p_{..k},\sum_{k=1}^{t}p_{..k}=1\right\}$$

A 中独立参数只有 $t\cdot(r+c-1)-1$ 个(习题 3.21).

该检验问题的似然比统计量为

$$\Lambda=\frac{\prod_{i=1}^{r}\prod_{j=1}^{c}\prod_{k=1}^{t}\left(\frac{n_{ijk}}{n}\right)^{n_{ijk}}}{\prod_{i=1}^{r}\prod_{j=1}^{c}\prod_{k=1}^{t}\left(\frac{n_{i,k}\cdot n_{.jk}}{n\cdot n_{..k}}\right)^{n_{ijk}}}$$

由于$(r\cdot c\cdot t-1)-[t\cdot(r+c-1)-1]=t\cdot(r-1)\cdot(c-1)$，故在原假设成立时，$2\ln\Lambda$ 的极限分布为 $\chi^2(t(r-1)(c-1))$. 在 $2\ln\Lambda\geqslant\chi^2_{1-\alpha}(t(r-1)(c-1))$时，即在

$$2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{t}n_{ijk}\cdot\ln\frac{n_{..k}\cdot n_{ijk}}{n_{i,k}\cdot n_{.jk}}\geqslant\chi^2_{1-\alpha}(t(r-1)(c-1))$$

的时候，拒绝原假设.

上述关于佛罗里达州的凶杀案件的例子中，考虑给定被害人的肤色后，被告的肤色与死刑判决的条件独立性的检验问题(给定被告的肤色后被害人的肤色与死刑判决的条件独立性的检验问题，以及给定是否死刑判决后被告的肤色与被害人的肤色的条件独立性的检验问题留作习题 3.22). 经计算，似然比检验统计量 $2\ln\Lambda=1.88$. 渐近 χ^2 分布的自由度为 2. 由于 $P\{\chi^2(2)\geqslant1.88\}\approx0.39$，所以我们不拒绝条件独立性原假设. 认为给定被害人的肤色后，被告的肤色与死刑判决是条件独立的. 由此可以看到，被害人的肤色与死刑判决有关，被害人的肤色与被告的肤色有关. 前者说明在美国有种族歧视.

与似然比检验法相类似的，还有 Wald 检验法和 Rao 检验法[3]. 此外，还有 Score 检验法[4].

§ 3.7 U 统计量检验

§ 3.7.1 U 统计量

上述各种检验方法基本上适用于参数统计结构. 这些方法往往要求总体分布族的密度函数的数学形式已知，且只含有限个未知参数. 但有的时候，人们难于由经验或某种理论得到总体的参数统计结构，而只能得到非参数统计结构. 因此有必要寻求非参数统计结构的检验方法.

首先看下面的例子.

例 3.17[5]　令 X 表示某种电子元器件的寿命. 根据经验或某种理论,常假设 X 服从指数分布,或 Weibull 分布,或对数正态分布等. 但有时很难作出这样的假设,或人们不愿意有这样的假设,而宁愿在一个较大的范围内考虑问题. 例如,仅假设 X 为取非负值的连续型随机变量. 这样我们就涉及到一个非参数统计结构.

如果对任意的 $s>0$,$t>0$,都有 $P(X>s+t|X>t)<P(X>s)$,则称该元件老化. 如果对任意的 $s>0$,$t>0$,都有 $P(X>s+t|X>t)=P(X>s)$,则称该元件无老化. 在可靠性统计中,一个很重要的检验问题是,

$$\text{原假设 } H_0\text{: 元件无老化} \quad \text{对} \quad \text{备择假设 } H_1\text{: 元件老化} \tag{3.68}$$

若 X 取参数统计结构,服从 Weibull 分布,其分布函数为 $F(x)=1-\exp\left\{-\frac{x^{\alpha}}{\sigma}\right\}$,则有

$$P(X>s)=\exp\left\{-\frac{s^{\alpha}}{\sigma}\right\}$$

$$P(X>s+t|X>t)=\frac{P(X>s+t,X>t)}{P(X>t)}=\frac{P(X>s+t)}{P(X>t)}$$

$$=\exp\left\{-\frac{(s+t)^{\alpha}-t^{\alpha}}{\sigma}\right\}$$

显然,在 $\alpha=1$ 时,$s^{\alpha}=(s+t)^{\alpha}-t^{\alpha}$,从而 $P(X>s+t|X>t)=P(X>s)$,元件无老化;而在 $\alpha>1$ 时,$s^{\alpha}<(s+t)^{\alpha}-t^{\alpha}$,从而 $P(X>s+t|X>t)<P(X>s)$,元件老化. 故上述检验问题在 Weibull 分布场合就简化为关于形状参数 α 的检验问题,原假设 H_0:$\alpha=1$ 对备择假设 H_1:$\alpha>1$. 受此启发,我们在 X 取非参数统计结构时,亦寻求一个参数,使得检验问题化为关于这个参数的检验问题.

设 X 为取非负值的连续随机变量,它的分布函数为 $F(x)$. 显然,

$$\text{“}P(X>s+t|X>t)=P(X>s)\text{”}\Leftrightarrow\text{“}\overline{F}(s+t)=\overline{F}(s)\cdot\overline{F}(t)\text{”}$$

$$\text{“}P(X>s+t|X>t)<P(X>s)\text{”}\Leftrightarrow\text{“}\overline{F}(s+t)<\overline{F}(s)\cdot\overline{F}(t)\text{”}$$

其中 $\overline{F}(x)=1-F(x)$. 为此,我们令

$$\omega = \int_0^\infty \int_0^\infty [\overline{F}(s) \cdot \overline{F}(t) - \overline{F}(s+t)] \mathrm{d}F(s) \mathrm{d}F(t)$$

则检验问题可简化为关于参数 ω 的检验问题

$$原假设\ H_0: \omega = 0 \quad 对 \quad 备择假设\ H_1: \omega > 0 \tag{3.69}$$

由于

$$\int_0^\infty \int_0^\infty \overline{F}(s) \cdot \overline{F}(t) \mathrm{d}F(s) \mathrm{d}F(t) = \int_0^1 \int_0^1 (1-u) \cdot (1-v) \mathrm{d}u \mathrm{d}v = \frac{1}{4}$$

$$\int_0^\infty \int_0^\infty \overline{F}(s+t) \mathrm{d}F(s) \mathrm{d}F(t) = \int_0^\infty \int_0^\infty P(X_1 > x_2 + x_3) \mathrm{d}F(x_2) \mathrm{d}F(x_3)$$
$$= P(X_1 > X_2 + X_3)$$

所以

$$\omega = \frac{1}{4} - P(X_1 > X_2 + X_3)$$

其中 X_1, X_2, X_3 是该元器件的寿命试验的任意三次记录.

若令 $\theta = P(X_1 > X_2 + X_3)$，则把检验问题(3.69)进一步简化为关于参数 θ 的检验问题

$$原假设\ H_0: \theta = \frac{1}{4} \quad 对 \quad 备择假设\ H_1: \theta < \frac{1}{4} \tag{3.70}$$

为了寻求检验统计量，首先得讨论 θ 的估计问题. 这是非参数估计问题. 设 $X = (X_1, \cdots, X_n)$ 是该元器件的寿命试验的 n 次记录，即它是来自取非负值的连续随机变量的分布函数 $F(x)$ 的样本. 容易看出

$$\psi(X_1, X_2, X_3) = \begin{cases} 1, & X_1 > X_2 + X_3 \\ 0, & X_1 \leqslant X_2 + X_3 \end{cases}$$

是 θ 的一个无偏估计. 考虑到样本中各个分量的地位是相同的，故定义对称函数

$$\phi(X_1, X_2, X_3) = \frac{\sum \psi(X_{i_1}, X_{i_2}, X_{i_3})}{3!}$$

它也是 θ 的一个无偏估计，其中 (i_1, i_2, i_3) 是 $(1, 2, 3)$ 的一个置换，Σ 对所有这种置换求和. $\psi(X_1, X_2, X_3)$ 和 $\phi(X_1, X_2, X_3)$ 都是 θ 的无偏估计，但 $\phi(X_1, X_2, X_3)$ 的方差不可能比 $\psi(X_1, X_2, X_3)$ 的方差大(习题 3.24).

进一步令

$$U(X_1,\cdots,X_n)=\frac{\sum\phi(X_{i_1},X_{i_2},X_{i_3})}{\binom{n}{3}} \tag{3.71}$$

其中 $1\leqslant i_1<i_2<i_3\leqslant n$，$\Sigma$ 对所有这种组合 (i_1,i_2,i_3) 求和. 则 $U(X_1,\cdots,X_n)$ 仍是 θ 的一个无偏估计. 它的方差不可能比 $\phi(X_1,X_2,X_3)$ 的方差大(习题 3.24). 由于在备择假设 H_1 为真时，θ 的值较小，故很自然地，我们在 $U(X_1,\cdots,X_n)\leqslant c$ 时，拒绝原假设，认为元件老化. 如何确定临界值 c 的问题留待下面解决.

形如(3.71)式的统计量就是第二章§2.2.4 给出的 U 统计量.

§3.7.2 U 统计量的期望和方差

设有样本 $X=(X_1,\cdots X_n)$，并设 ϕ 为 $X_1,\cdots,X_m(m\leqslant n)$ 的对称函数，令

$$U=U(X_1,\cdots,X_n)=\binom{n}{m}^{-1}\cdot\sum\phi(X_{i_1},\cdots,X_{i_m}) \tag{3.72}$$

其中 $1\leqslant i_1<\cdots<i_m\leqslant n$，$\Sigma$ 对所有的组合 $(i_1,\cdots,i_m)$ 求和. 则 U(或 $U(X_1,\cdots,X_n)$)是以 $\phi(X_1,\cdots,X_m)$ 为核的 U 统计量.

若令 $\theta=\mathrm{E}[\phi(X_1,\cdots,X_m)]$，则 U 统计量的数学期望 $\mathrm{E}[U]=\theta$. 为简化 U 统计量的方差的计算，不妨假设 $\theta=0$. 否则，以 $\phi(X_1,\cdots,X_m)-\theta$ 为核构造 U 统计量. 下面，我们来计算 U 的方差. 为此首先在 $X_1,\cdots,X_c(c\leqslant m)$ 给定的条件下，计算 $\phi(X_1,\cdots,X_m)$ 的条件期望

$$\phi_c(x_1,\cdots,x_c)=\mathrm{E}[\phi(X_1,\cdots,X_m)\mid X_1=x_1,\cdots,X_c=x_c] \tag{3.73}$$

其中 $c=1,\cdots,m$. 则

$$\begin{aligned}\mathrm{E}[\phi_c(X_1,\cdots,X_c)]&=\mathrm{E}\{\mathrm{E}[\phi(X_1,\cdots,X_m)\mid X_1,\cdots,X_c]\}\\&=\mathrm{E}[\phi(X_1,\cdots,X_m)]=0\end{aligned}$$

记

$$\mathrm{E}[\phi_c(X_1,\cdots,X_c)]^2=\sigma_c^2,\quad c=1,\cdots,m \tag{3.74}$$

若 $\phi(X_1,\cdots,X_m)$ 的方差有限，则 σ_c^2 也有限(习题 3.24). 由(3.72)式知，

U 统计量的方差为

$$\mathrm{Var}(U)=\binom{n}{m}^{-2}\cdot\sum \mathrm{E}\left[\phi(X_{i_1},\cdots,X_{i_m})\cdot\phi(X_{j_1},\cdots,X_{j_m})\right]$$

其中 $1\leqslant i_1<\cdots<i_m\leqslant n$，$1\leqslant j_1<\cdots<j_m\leqslant n$，$\Sigma$ 对所有的组合 $(i_1,\cdots,i_m)$ 和 $(j_1,\cdots,j_m)$ 求和. 显然，在两集合 $\{i_1,\cdots,i_m\}$ 和 $\{j_1,\cdots,j_m\}$ 无公共元素时，有

$$\mathrm{E}\left[\phi(X_{i_1},\cdots,X_{i_m})\cdot\phi(X_{j_1},\cdots,X_{j_m})\right]=0$$

所以

$$\mathrm{Var}(U)=\binom{n}{m}^{-2}\cdot\sum_{c=1}^{m}M(c)$$

其中

$$M(c)=\sum \mathrm{E}\left[\phi(X_{i_1},\cdots,X_{i_m})\cdot\phi(X_{j_1},\cdots,X_{j_m})\right] \qquad (3.75)$$

这里 $1\leqslant i_1<\cdots<i_m\leqslant n$，$1\leqslant j_1<\cdots<j_m\leqslant n$，且两集合 $\{i_1,\cdots,i_m\}$ 和 $\{j_1,\cdots,j_m\}$ 中恰有 $c(c=1,\cdots,m)$ 个元素相同，Σ 对所有这样的组合 $(i_1,\cdots,i_m)$ 和 $(j_1,\cdots,j_m)$ 求和.

由样本的独立同分布性知，在两集合 $\{i_1,\cdots,i_m\}$ 和 $\{j_1,\cdots,j_m\}$ 中恰有 c 个元素相同时，有

$$\begin{aligned}&\mathrm{E}\left[\phi(X_{i_1},\cdots,X_{i_m})\cdot\phi(X_{j_1},\cdots,X_{j_m})\right]\\&=\mathrm{E}\left[\phi(X_1,\cdots,X_m)\cdot\phi(X_1,\cdots,X_c,X_{m+1},\cdots,X_{2m-c})\right]\end{aligned}$$

所以(3.75)右边和式中的各个项都相等. 因此

$$\begin{aligned}M(c)=&\binom{n}{m}\cdot\binom{m}{c}\cdot\binom{n-m}{m-c}\cdot\\&\mathrm{E}\left[\phi(X_1,\cdots,X_m)\cdot\phi(X_1,\cdots,X_c,X_{m+1},\cdots,X_{2m-c})\right]\end{aligned}$$

在 $X_1=x_1,\cdots,X_c=x_c$ 给定的条件下，计算 $\phi(X_1,\cdots,X_m)\cdot\phi(X_1,\cdots,X_c,X_{m+1},\cdots,X_{2m-c})$ 的条件期望. 由于在 $X_1=x_1,\cdots,X_c=x_c$ 给定时，$(X_1,\cdots,X_m)$ 和 $(X_1,\cdots,X_c,X_{m+1},\cdots,X_{2m-c})$ 相互条件独立和同条件分布，则有

$$\begin{aligned}&\mathrm{E}\left[\phi(X_1,\cdots,X_m)\cdot\phi(X_1,\cdots,X_c,X_{m+1},\cdots,X_{2m-c})\,|\,X_1=x_1,\cdots,X_c=x_c\right]\\&=\mathrm{E}\left[\phi(X_1,\cdots,X_m)\,|\,X_1=x_1,\cdots,X_c=x_c\right]\times\end{aligned}$$

$$\mathrm{E}[\phi(X_1,\cdots,X_c,X_{m+1},\cdots,X_{2m-c})|X_1=x_1,\cdots,X_c=x_c]$$
$$=[\phi_c(x_1,\cdots,x_c)]^2$$

所以

$$\mathrm{E}[\phi(X_1,\cdots,X_m)\cdot\phi(X_1,\cdots,X_c,X_{m+1},\cdots,X_{2m-c})]$$
$$=\mathrm{E}\{\mathrm{E}[\phi(X_1,\cdots,X_m)\cdot\phi(X_1,\cdots,X_c,X_{m+1},\cdots,X_{2m-c})|X_1,\cdots,X_c]\}$$
$$=\mathrm{E}\{[\phi_c(X_1,\cdots,X_c)]^2\}=\sigma_c^2$$

从而 $M(c)=\begin{pmatrix}n\\m\end{pmatrix}\cdot\begin{pmatrix}m\\c\end{pmatrix}\cdot\begin{pmatrix}n-m\\m-c\end{pmatrix}\cdot\sigma_c^2$. 故 U 统计量的方差为

$$\mathrm{Var}[U]=\begin{pmatrix}n\\m\end{pmatrix}^{-1}\cdot\sum_{c=1}^{m}\begin{pmatrix}m\\c\end{pmatrix}\cdot\begin{pmatrix}n-m\\m-c\end{pmatrix}\cdot\sigma_c^2$$
$$=\sum_{c=1}^{m}\begin{pmatrix}m\\c\end{pmatrix}\cdot\frac{m!}{(m-c)!}\cdot$$
$$\frac{(n-m)(n-m-1)\cdots(n-2m+c+1)}{n(n-1)\cdots(n-m+1)}\cdot\sigma_c^2$$

由于 $(n-m)(n-m-1)\cdots(n-2m+c+1)$ 和 $n(n-1)\cdots(n-m+1)$ 分别是 n 的 $m-c$ 和 m 次多项式，所以

$$\mathrm{Var}[U]=\sum_{c=1}^{m}\begin{pmatrix}m\\c\end{pmatrix}\cdot\frac{m!}{(m-c)!}\cdot[n^{-c}+O(n^{-c-1})]\cdot\sigma_c^2$$
$$=m^2\cdot[n^{-1}+O(n^{-2})]\sigma_1^2+\sum_{c=2}^{m}\begin{pmatrix}m\\c\end{pmatrix}\cdot\frac{m!}{(m-c)!}\cdot$$
$$[n^{-c}+O(n^{-c-1})]\cdot\sigma_c^2$$
$$=\frac{m^2}{n}\cdot\sigma_1^2+O(n^{-2})\tag{3.76}$$

在例 3.17 中，$m=3$. 当原假设 H_0 成立时，$\theta=\frac{1}{4}$，则 $\mathrm{E}[U]=\theta=\frac{1}{4}$. 由(3.73)式及样本的独立同分布性知

$$\phi_1(x_1)=\mathrm{E}[\phi(X_1,X_2,X_3)|X_1=x_1]$$
$$=\frac{1}{3}\cdot P(x_1>X_2+X_3)+\frac{2}{3}\cdot P(X_2>X_3+x_1)$$
$$=\frac{1}{3}\cdot\int_0^{x_1}F(x_1-x_3)\mathrm{d}F(x_3)+\frac{2}{3}\cdot\int_0^{\infty}\overline{F}(x_3+x_1)\mathrm{d}F(x_3)$$

当原假设成立时，由于对任意的 $s>0, t>0$，都有 $\overline{F}(s+t)=\overline{F}(s)\cdot\overline{F}(t)$，则有

$$F(x_1 - x_3) = 1 - \frac{\overline{F}(x_1)}{\overline{F}(x_3)}, \quad \overline{F}(x_3 + x_1) = \overline{F}(x_3)\cdot\overline{F}(x_1)$$

从而

$$\begin{aligned}\phi_1(x_1) &= \frac{1}{3}\cdot\int_0^{x_1}\left[1 - \frac{\overline{F}(x_1)}{\overline{F}(x_3)}\right]\mathrm{d}F(x_3) + \frac{2}{3}\cdot\int_0^{\infty}F(x_3)\overline{F}(x_1)\mathrm{d}F(x_3)\\ &= \frac{1}{3}\cdot[F(x_1) + \overline{F}(x_1)\cdot\ln\overline{F}(x_1)] + \frac{1}{3}\cdot\overline{F}(x_1)\\ &= \frac{1}{3}\cdot[1 + \overline{F}(x_1)\cdot\ln\overline{F}(x_1)],\end{aligned}$$

下面计算 $\phi_1(X_1)$ 的方差 σ_1^2. 由于当原假设成立时，

$$\mathrm{E}[\phi_1(X_1)] = \frac{1}{4}$$

$$\begin{aligned}\mathrm{E}[\phi_1(X_1)]^2 &= \frac{1}{9}\cdot\int_0^{\infty}[1 + \overline{F}(x_1)\cdot\ln\overline{F}(x_1)]^2\mathrm{d}F(x_1)\\ &= \frac{1}{9}\int_0^1[1 + (1-u)\cdot\ln(1-u)]^2\mathrm{d}u = \frac{31}{486}\end{aligned}$$

所以

$$\sigma_1^2 = \frac{31}{486} - \left(\frac{1}{4}\right)^2 = \frac{5}{3888}$$

同理（习题 3.25），在原假设成立时，

$$\phi_2(x_1, x_2) = \begin{cases}[F(x_1 - x_2) + \overline{F}(x_1 + x_2)]/3, & x_1 > x_2\\ [F(x_2 - x_1) + \overline{F}(x_1 + x_2)]/3, & x_2 > x_1\end{cases}$$

$$\sigma_2^2 = \frac{7}{1296}$$

$$\phi_3(x_1, x_2, x_3) = \phi(x_1, x_2, x_3)$$

$$\sigma_3^2 = \frac{1}{48}$$

所以

$$\mathrm{Var}[U] = \frac{(n-3)(n-4)}{n(n-1)(n-2)}\cdot\frac{5}{432} + \frac{n-3}{n(n-1)(n-2)}\cdot\frac{7}{72} +$$

$$\frac{1}{n(n-1)(n-2)} \cdot \frac{1}{8}$$

$$= \frac{5}{432n} + O(n^{-2})$$

§ 3.7.3　U 统计量的渐近正态性

定理 3.20　对由(3.72)式定义的以 $\phi(X_1, \cdots, X_m)$ 为核的 U 统计量，在核函数 $\phi(X_1, \cdots, X_m)$ 的数学期望为 θ 且方差有限时，

$$\sqrt{n} \cdot (U-\theta) \xrightarrow{L} N(0, m^2\sigma_1^2), \quad n \to \infty$$

其中 $\sigma_1^2 > 0$，$\sigma_c^2 (c=1, \cdots, m)$ 由(3.73)和(3.74)两式给出.

证明：不妨假设 $\theta=0$. 否则以 $\phi(X_1, \cdots, X_m)-\theta$ 代替 $\phi(X_1, \cdots, X_m)$.

显然，$\phi_1(X_1), \cdots, \phi_1(x_n)$ 独立同分布，其中函数 ϕ_1 由(3.73)式给出. 令 $V=\frac{m}{n} \cdot \sum_{i=1}^{n} \phi_1(X_i)$，则由中心极限定理知，

$$\sqrt{n} \cdot V \xrightarrow{L} N(0, m^2\sigma_1^2), \quad n \to \infty$$

因而，为证本定理，只需证明 $n \to \infty$ 时，$\sqrt{n} \cdot (U-V) \xrightarrow{p} 0$. 由于 $\mathrm{E}\{\sqrt{n} \cdot (U-V)\}=0$，故只需证明 $n \to \infty$ 时，$\mathrm{Var}\{\sqrt{n} \cdot (U-V)\} \to 0$. 由(3.76)式知，

$$\mathrm{Var}\{\sqrt{n} U\}=n \cdot \left[\frac{m^2}{n} \cdot \sigma_1^2+O(n^{-2})\right] \to m^2\sigma_1^2,$$

$$n \to \infty \tag{3.77}$$

由于 $\phi_1(X_1)$ 的方差为 σ_1^2，所以

$$\mathrm{Var}\{\sqrt{n} V\}=m^2\sigma_1^2 \tag{3.78}$$

下面计算 $\mathrm{Cov}(\sqrt{n} \cdot U, \sqrt{n} \cdot V)$.

$$\mathrm{Cov}(\sqrt{n} \cdot U, \sqrt{n} \cdot V)=\mathrm{E}\{U \cdot nV\}=m \cdot \sum_{i=1}^{n} \mathrm{E}\{U \cdot \phi_1(X_i)\}$$

$$=mn \cdot \mathrm{E}\{U \cdot \phi_1(X_1)\}$$

$$=mn \cdot \binom{n}{m}^{-1} \cdot \sum \mathrm{E}\{\phi(X_{i_1}, \cdots, X_{i_m}) \cdot \phi_1(X_1)\}$$

其中 $1\leqslant i_1<\cdots<i_m\leqslant n$，$\Sigma$ 对所有的组合$(i_1,\cdots,i_m)$求和. 若 $i_1>1$，则 $\mathrm{E}\{\phi(X_{i_1},\cdots,X_{i_m})\cdot\phi_1(X_1)\}=0$. 从而

$$
\begin{aligned}
&\mathrm{Cov}(\sqrt{n}\cdot U,\sqrt{n}\cdot V)\\
&=mn\cdot\binom{n}{m}^{-1}\cdot\sum\mathrm{E}\{\phi(X_1,X_{i_2},\cdots,X_{i_m})\cdot\phi_1(X_1)\}
\end{aligned}
$$

其中 $1<i_2<\cdots<i_m\leqslant n$，$\Sigma$ 对所有的组合$(i_2,\cdots,i_m)$求和. 因此

$$
\begin{aligned}
&\mathrm{Cov}(\sqrt{n}\cdot U,\sqrt{n}\cdot V)\\
&=mn\cdot\binom{n}{m}^{-1}\cdot\binom{n-1}{m-1}\cdot\mathrm{E}\{\phi(X_1,\cdots,X_m)\cdot\phi_1(X_1)\}
\end{aligned}
$$

在 $X_1=x_1$ 给定的条件下，计算 $\phi(X_1,\cdots,X_m)\cdot\phi_1(X_1)$的条件期望

$$
\mathrm{E}\{\phi(X_1,\cdots,X_m)\cdot\phi_1(X_1)\mid X_1=x_1\}=[\phi_1(X_1)]^2
$$

则

$$
\mathrm{Cov}(\sqrt{n}\cdot U,\sqrt{n}\cdot V)=mn\cdot\binom{n}{m}^{-1}\cdot\binom{n-1}{m-1}\cdot\sigma_1^2=m^2\sigma_1^2 \tag{3.79}
$$

从而由(3.77)，(3.78)和(3.79)三式知，在 $n\to\infty$时，

$$
\begin{aligned}
&\mathrm{Var}\{\sqrt{n}\cdot(U-V)\}\\
&=\mathrm{Var}\{\sqrt{n}U\}+\mathrm{Var}\{\sqrt{n}V\}-\\
&\quad 2\cdot\mathrm{Cov}(\sqrt{n}\cdot U,\sqrt{n}\cdot V)\to 0
\end{aligned}
$$

定理证毕.

在例 3.17，$m=3$，$\sigma_1^2=\frac{5}{3888}$，在原假设 H_0 成立时，

$$
\sqrt{n}\cdot\left(U-\frac{1}{4}\right)\xrightarrow{L}N\left(0,\frac{5}{432}\right)
$$

故在大样本场合下检验问题(3.68)的拒绝域为

$$
U\leqslant\frac{1}{4}-\sqrt{\frac{5}{432n}}\cdot U_{1-\alpha}
$$

其中 $U_{1-\alpha}$为标准正态分布 $N(0,1)$的 $1-\alpha$ 分位数.

例 3.18(对称中心的检验)　设总体 X 的分布函数为 $F(x-\theta)$，其

中 $F(x)$ 是关于原点对称的连续型分布. 由此得到了一个非参数统计结构. 考虑检验问题

$$原假设\ H_0: \theta = 0 \quad 对 \quad 备择假设\ H_1: \theta > 0 \tag{3.80}$$

即检验对称分布的对称中心在原点，还是在原点的右边. 设 $X_1, \cdots, X_n$ 是来自总体 X 的一个样本. 若 X 取参数统计结构，例如服从正态分布 $N(\mu, \sigma^2)$，那么上述检验问题就是关于均值 μ 的检验问题. 常用 t 检验法. 在 X 取非参数统计结构时，注意到

"原假设 $H_0: \theta = 0$ 成立"⇒"$P\{X_1 + X_2 > 0\} = 0.5$"；

"备择假设 $H_1: \theta > 0$ 成立"⇒"$P\{X_1 + X_2 > 0\} > 0.5$".

因而令 $\omega = P\{X_1 - X_2 > 0\}$. 那么，上述检验问题可简化为关于参数 ω 的检验问题.

$$原假设\ H_0: \omega = 0.5\ 对备择假设\ H_1: \omega > 0.5 \tag{3.81}$$

显然，

$$\phi(X_1, X_2) = \begin{cases} 1, & X_1 + X_2 > 0 \\ 0, & X_1 + X_2 \leqslant 0 \end{cases}$$

是 ω 的一个无偏估计. 然后由对称函数 $\phi(X_1, X_2)$ 构造 U 统计量，

$$U = U(X_1, \cdots, X_n) = \frac{\sum \phi(X_{i_1}, X_{i_2})}{\binom{n}{2}}$$

其中 $1 \leqslant i_1 < i_2 \leqslant n$，$\Sigma$ 对所有的组合 (i_1, i_2) 求和. U 也是 ω 的一个无偏估计. 由于在备择假设成立时，ω 的值较大，所以我们在 $U(X_1, \cdots, X_n) \geqslant c$ 时，拒绝原假设.

这里 $m = 2$. 在原假设 H_0 为真时，$\mathrm{E}[U(X_1, \cdots, X_n)] = 0.5$，并且不难推得(习题 3.26)，

$$\phi_1(x_1) = F(x_1), \sigma_1^2 = \frac{1}{12}$$

$$\phi_2(x_1, x_2) = \phi(x_1, x_2), \sigma_2^2 = \frac{1}{4}$$

所以

$$\mathrm{Var}[U(X_1, \cdots, X_n)] = \frac{2n-1}{6n(n-1)} = \frac{1}{3n} + O(n^{-2})$$

在原假设成立时，$\sqrt{n}\cdot\left(U-\frac{1}{2}\right)\xrightarrow{L}N\left(0,\frac{1}{3}\right),n\to\infty$. 故在大样本场合下检验问题(3.80)的拒绝域为

$$U\geqslant\frac{1}{2}+\sqrt{\frac{1}{3n}}\cdot U_{1-\alpha}$$

§ 3.7.4 两样本 U 统计量

定义 3.12 样本 $X_1,\cdots,X_{n_1}$ 和 $Y_1,\cdots,Y_{n_2}$ 分别来自相互独立的总体 X 和 Y. 设在 $X_1,\cdots,X_{m_1}$ 给定时，$\phi(X_1,\cdots,X_{m_1};Y_1,\cdots,Y_{m_2})$ 为 $Y_1,\cdots,Y_{m_2}$ 的对称函数，而在 $Y_1,\cdots,Y_{m_2}$ 给定时，

$$\phi(X_1,\cdots,X_{m_1};Y_1,\cdots,Y_{m_2})$$

为 $X_1,\cdots,X_{m_1}$ 的对称函数. 令

$$\begin{aligned}U&=U(X_1,\cdots,X_{n_1};Y_1,\cdots,Y_{n_2})\\&=\binom{n_1}{m_1}^{-1}\cdot\binom{n_2}{m_2}^{-1}\cdot\sum\phi(X_{i_1},\cdots,X_{i_{m_1}};Y_{j_1},\cdots,Y_{j_{m_2}})\end{aligned}\tag{3.82}$$

其中 $1\leqslant i_1<\cdots<i_{m_1}\leqslant n_1$，$1<j_1<\cdots<j_{m_2}\leqslant n_2$，$\sum$ 是对所有的组合 $(i_1,\cdots,i_{m_1})$ 和 $(j_1,\cdots,j_{m_2})$ 求和. 则称 U 是以 $\phi(X_1,\cdots,X_{m_1};Y_1,\cdots,Y_{m_2})$ 为核的两样本 U 统计量.

设 $\mathrm{E}[\phi(X_1,\cdots,X_{m_1};Y_1,\cdots,Y_{m_2})]=\theta$，则 U 统计量的数学期望

$$\mathrm{E}[U]=\theta$$

与单样本的情况相类似地，令

$$\begin{aligned}&\phi_{c_1c_2}(x_1,\cdots,x_{c_1};y_1,\cdots,y_{c_2})\\&=\mathrm{E}[\phi(X_1,\cdots,X_{m_1};Y_1,\cdots,Y_{m_2})\mid X_1=x_1,\cdots,X_{c_1}=x_{c_1};\\&\quad Y_1=y_1,\cdots,Y_{c_2}=y_{c_2}]\\&\sigma^2_{c_1c_2}=\mathrm{Var}[\phi_{c_1c_2}(X_1,\cdots,X_{c_1};Y_1,\cdots,Y_{c_2})]\end{aligned}\tag{3.83}$$

其中 $c_1=0,1,\cdots,m_1$；$c_2=0,1,\cdots,m_2$. 注意：在 $c_1=c_2=0$ 时，

$$\phi_{00}(x_1,\cdots,x_{c_1};y_1,\cdots,y_{c_2})=\mathrm{E}[\phi(X_1,\cdots,X_{m_1};Y_1,\cdots,Y_{m_2})]=\theta$$

$$\sigma_{00}=0$$

若 $\phi(X_1,\cdots,X_{m_1};Y_1,\cdots,Y_{m_2})$ 的方差有限，则 $\sigma^2_{c_1c_2}$ 也有限. 与单样本的

情况相类似地，可以得到 U 统计量的方差，它等于

$$\operatorname{Var}[U]=\binom{n_1}{m_1}^{-1}\cdot\binom{n_2}{m_2}^{-1}\cdot\sum_{c_1=0}^{m_1}\sum_{c_2=0}^{m_2}\binom{m_1}{c_1}\cdot$$

$$\binom{n_1-m_1}{m_1-c_1}\cdot\binom{m_2}{c_2}\cdot\binom{n_2-m_2}{m_2-c_2}\cdot\sigma_{c_1c_2}^2$$

$$=\frac{m_1^2}{n_1}\cdot\sigma_{10}^2+\frac{m_2^2}{n_2}\cdot\sigma_{01}^2+O(n_1^{-2}+n_2^{-2})$$

定理 3.21　对由(3.82)式定义的以 $\phi(X_1,\cdots,X_{m_1};Y_1,\cdots,Y_{m_2})$ 为核的两样本 U 统计量，在核函数 $\phi(X_1,\cdots,X_{m_1};Y_1,\cdots,Y_{m_2})$ 的数学期望为 θ 且方差有限时，

$$\frac{U-\theta}{\sqrt{\frac{m_1^2\sigma_{10}^2}{n_1}+\frac{m_2^2\sigma_{01}^2}{n_2}}}\xrightarrow{L}N(0,1),\quad n_1\to\infty,\quad n_2\to\infty$$

其中 $\sigma_{10}^2>0,\sigma_{01}^2>0,\sigma_{c_1c_2}^2(c_1=0,1,\cdots,m_1;c_2=0,1,\cdots,m_2)$ 由(3.83)式给出.

证明与单样本的情况相类似[5].

例 3.19（位置参数的 Mann-Whitney 检验）　样本 $X_1,\cdots,X_{n_1}$ 和 $Y_1,\cdots,Y_{n_2}$ 分别来自相互独立的连续随机变量总体 X 和 Y. 设 X 和 Y 的分布函数分别为 F 和 G. 考虑检验问题，原假设 H_0：对任意的 x，都有 $F(x)=G(x)$，备择假设 H_1：对任意的 x，都有 $F(x)>G(x)$.

在 X 和 Y 服从方差相等的正态分布时，这就是比较两个正态分布的均值的检验问题，常用 t 检验法.

在 $G(x)=F(x-a)$ 时，这就是关于位置参数 a 的检验问题，原假设即变为 H_0：$a=0$，而备择假设即变为 H_1：$a>0$. 这个检验问题常称为位置参数的检验问题，受人关注.

在原假设成立，即对任意的 x，都有 $F(x)=G(x)$ 时，

$$P(X<Y)=\int_{-\infty}^{\infty}F(y)\mathrm{d}G(y)=\int_0^1u\mathrm{d}u=0.5\qquad(3.84)$$

而在备择假设为真，即对任意的 x，都有 $F(x)>G(x)$ 时，

$$P(X<Y)=\int_{-\infty}^{\infty}F(y)\mathrm{d}G(y)>\int_{-\infty}^{\infty}G(y)\mathrm{d}G(y)$$

$$= \int_0^1 u\mathrm{d}u = 0.5 \tag{3.85}$$

在 $P(X<Y)>0.5$ 时，这时我们称 X 随机地比 Y 小，记为"$X\overset{s}{<}Y$". 由(3.84)和(3.85)两式知，我们令 $\omega=P(X<Y)$. 那么上述检验问题可简化为关于参数 ω 的检验问题，原假设 H_0：$\omega=0.5$ 对备择假设 H_1：$\omega>0.5$. 显然，以

$$\phi(X_1,Y_1) = \begin{cases} 1, & X_1 < Y_1 \\ 0, & X_1 \geqslant Y_1 \end{cases}$$

为核构造的 U 统计量，

$$\begin{aligned} U &= U(X_1,\cdots,X_{n_1};Y_1,\cdots,Y_{n_2}) \\ &= n_1^{-1}\cdot n_2^{-1}\cdot \sum \phi(X_i;Y_j) \end{aligned} \tag{3.86}$$

是 ω 的无偏估计，其中 $1\leqslant i\leqslant n_1, 1\leqslant j\leqslant n_2$. $\sum$ 对所有的 i 和 j 求和. 该 U 统计量是 Mann 和 Whitney 在 1947 年提出的，人们称它为 Mann-Whitney 统计量. 由于在备择假设为真时，ω 的值较大，所以我们在 $U\geqslant c$ 时，拒绝原假设，认为"对任意的 x，都有 $F(x)>G(x)$".

在原假设成立时，$\omega=0.5$，所以 $\mathrm{E}[U]=0.5$. 同时还有(习题3.27)

$$\phi_{10}(x_1) = G(x_1) = F(x_1), \quad \phi_{01}(y_1) = \overline{F}(y_1)$$

$$\phi_{11}(x_1,y_1) = \phi(x_1,y_1)$$

$$\sigma_{10}^2 = \sigma_{01}^2 = \frac{1}{12}, \quad \sigma_{11}^2 = \frac{1}{4}$$

所以

$$\mathrm{Var}[U] = \frac{n_1+n_2+1}{12n_1n_2}$$

据定理 3.21 知，在 $n_1\to\infty, n_2\to\infty$ 时，

$$\sqrt{12n_1n_2}\cdot\frac{U-0.5}{\sqrt{n_1+n_2}} \xrightarrow{L} N(0,1) \tag{3.87}$$

故在大样本场合下检验的拒绝域为

$$U \geqslant \frac{1}{2} + \sqrt{\frac{n_1+n_2}{12n_1n_2}}\cdot U_{1-\alpha}$$

这个检验称为Mann-Whitney 检验，或位置参数的 Mann-Whitney 检验.

参考文献

1 Lehmann E L. Testing Statistical Hypotheses (second Edition). John Wiley & Sons, New York, 1986

2 陈希孺. 数理统计引论. 北京：科学出版社，1981

3 Rao C R. Linear Statistical Inference and Its Applications. John Wiley & Sons, New York, 1973

4 Cox D R and Hinkley D V. Theoretical Statistics. Chapman and Hall, London. 1974

5 陈希孺等. 非参数统计. 上海：上海科学技术出版社，1989

6 Agresti A. Analysis of Ordinal Categorical Data. John Wiley & Sons, New York, 1984

习 题 三

3.1　电话交换台单位时间内接到的呼唤次数服从泊松分布 $P(\lambda)$，$\lambda>0$. λ为单位时间内接到的平均呼唤次数. 设 $x=(x_1,\cdots,x_{10})$是该电话交换台的 10 次记录. 考虑假设检验问题：原假设 $H_0:\lambda\geqslant 1$ 对备择假设 $H_1:\lambda<1$. 取水平为 $\alpha=0.05$.

3.2　设 $X=(X_1,\cdots,X_n)$是来自正态分布族$\{N(0,\sigma^2):0<\sigma^2<\infty\}$的样本，考虑原假设 $H_0:\sigma^2=1$ 对备择假设 $H_1:\sigma^2=\sigma_1^2(\sigma_1^2>1)$的检验问题. 取水平为 $\alpha(0<\alpha<1)$. 试求其 MPT.

3.3　设 $X=(X_1,\cdots,X_n)$是来自均匀分布族$\{R(0,\theta):\theta>0\}$的样本，考虑原假设 $H_0:\theta=1$ 对备择假设 $H_1:\theta=\theta_1(\theta_1<1)$的检验问题. 取水平为 $\alpha(0<\alpha<1)$. 试求其 MPT.

3.4　(N-P 基本引理的补遗)设 P_{θ_0}和 P_{θ_1}是可测空间$(\mathscr{X},\mathscr{B})$上的两个不同的概率测度，关于某个 σ 有限的测度 μ，有

$$p(x;\theta_0)=\frac{\mathrm{d}P_{\theta_0}}{\mathrm{d}\mu},\quad p(x;\theta_1)=\frac{\mathrm{d}P_{\theta_1}}{\mathrm{d}\mu}$$

则在检验问题(3.9)中，如果 $\phi(x)$ 是水平为 α 的 MPT，则它不一定满足(3.10)式，但在 $E_{\theta_1}\phi(X)<1$ 的时候，它必满足(3.10)式.

3.5 设样本 $X=(X_1,\cdots,X_{100})$ 来自二点分布族 $\{b(1,p):0\leqslant p\leqslant 1\}$. 试求检验问题：$H_0:p\leqslant 0.01$ 对 $H_1:p>0.01$ 的水平 $\alpha=0.05$ 的 UMPT.

3.6 (定理 3.7 的推广)设概率密度族 $\{p(x;\theta):\theta\in\Theta\subseteq\mathbf{R}\}$ 关于 X 具有非降 MLR. 试证明：

(1)若 $X_1,\cdots,X_n$ 是来自该 MLR 分布族的一个样本，并且 $\psi(x_1,\cdots,x_n)$ 关于每一个 x_i 都是非降的，则 $E_\theta\psi(X_1,\cdots,X_n)$ 是 θ 的一个非降函数；

(2)若函数 $\psi(x)$ 具有性质：存在点 x_0 使得在 $x<x_0$ 时，$\psi(x)\leqslant 0$，而在 $x>x_0$ 时，$\psi(x)\geqslant 0$，则 $E_\theta\psi(X)$ 或者总是正的，或者总是负的，或者存在点 θ_0 使得在 $\theta<\theta_0$ 时，$E_\theta\psi(X)\leqslant 0$，而在 $\theta>\theta_0$ 时，$E_\theta\psi(X)\geqslant 0$.

3.7 设 X 和 Y 的密度函数分别为 $f(x)$ 和 $g(y)$. 试证明：若 $f(x)/g(x)$ 是 x 的非降函数，则 $E(X)\geqslant E(Y)$.

3.8 设 $X=(X_1,\cdots,X_n)$ 是来自带有位置参数的指数分布总体的样本. 总体的密度函数为

$$p(x;\theta)=\begin{cases}\exp\{-(x-\theta)\}, & x\geqslant\theta\\ 0, & \text{其它}\end{cases}$$

考虑如下的检验问题：原假设 $H_0:\theta=0$ 对备择假设 $H_1:\theta>0$. 试构造水平为 $\alpha(0<\alpha<1)$ 的 UMPT.

3.9 设 $X=(X_1,\cdots,X_{10})$ 是来自 Pareto 分布总体的样本. Pareto 分布的密度函数为

$$p(x;\theta)=\begin{cases}\dfrac{\beta}{\theta}\cdot\left(\dfrac{\theta}{x}\right)^{\beta+1}, & x\geqslant\theta\\ 0, & \text{其它}\end{cases}$$

其中 $\beta=2$ 已知. 考虑如下的检验问题：原假设 $H_0:\theta=1$ 对备择假设 $H_1:\theta\neq 1$. 试构造水平 $\alpha=0.1$ 的 UMPT.

3.10 设 $T(x)$ 的密度函数如(3.26)式所示，即为

$$p(t;\theta)=c(\theta)\cdot\exp\{\theta\cdot t\}\cdot h(t)$$

试证明：$E_\theta[T]=-\dfrac{c'(\theta)}{c(\theta)}$.

3.11　设 $T(x)$ 的密度函数如(3.26)式所示，其中 $c(\theta)>0$. 单边假设检验问题：原假设 $H_0:\theta\leqslant\theta_0$ 对备择假设 $H_1:\theta>\theta_0$ 的水平为 $\alpha(0<\alpha<1)$ 的 UMP 检验 $\phi(T)$ 如(3.19)所示，即为

$$\phi(T)=\begin{cases}1, & T>c\\ r, & T=c\\ 0, & T<c\end{cases}$$

其中常数 $r(0\leqslant r\leqslant 1)$ 和 c 由(3.20)式，即由 $E_{\theta_0}\phi(T)=\alpha$ 确定. 我们知道这个检验的势函数 $g(\theta)=E_\theta\phi(T(X))$ 是非降的，且在集合 $\{\theta:0<g(\theta)<1\}$ 上是严格增加的. 试证明：$g'(\theta_0)>0$.

提示：为了证明 $g'(\theta_0)>0$，考虑如下的问题：在满足条件 $E_{\theta_0}\phi(T)=\alpha$ 的检验函数 $\phi(T)$ 中寻找一个检验，使得其势函数在点 θ_0 处的导数 $g'(\theta_0)$ 达到最大，或等价地使得 $E_{\theta_0}[\phi(T)\cdot T]$ 最大.

3.12　设样本 X 服从多参数指数型分布(3.47). 试证明：原假设 $H_0:\theta=\theta_0$ 对备择假设 $H_1:\theta\neq\theta_0$ 的双边假设检验问题(Ⅲ)的水平为 α 的 UMPUT 由(3.53)式给出，其中 $r_i(t)(0\leqslant r_i(t)\leqslant 1)$ 和 $c_i(t)$ 由(3.50)和(3.53)两式确定.

3.13　设样本 $X=(X_1,\cdots,X_n)$ 来自正态总体 $N(\mu,\sigma^2)$，其中 $-\infty<\mu<\infty,\sigma^2>0$. 在 $\mu=0$ 时，试证明 w 的分布与参数 σ^2 无关，且 w 的分布关于原点对称，其中

$$w=\frac{\overline{X}}{\sqrt{\sum_{i=1}^{n}X_i^2}}$$

3.14　设样本 $X=(X_1,\cdots,X_m)$ 和样本 $Y=(Y_1,\cdots,Y_n)$ 互相独立. 它们分别来自正态总体 $N(\mu_1,\sigma^2)$ 和 $N(\mu_2,\sigma^2)$，其中：$-\infty<\mu_1<\infty$，$-\infty<\mu_2<\infty,\sigma^2>0$. 在 $\mu_1=\mu_2=\mu$ 时，试证明：V 的分布与参数 σ^2 和 μ 无关，V 的分布关于原点对称，其中

$$V=\frac{U}{\sqrt{T_2-\dfrac{T_1^2}{m+n}}},\quad U=\overline{Y}-\overline{X}$$

$$T_1 = m\overline{X} + n\overline{Y}, T_2 = \sum_{i=1}^{m} X_i^2 - \sum_{i=1}^{n} Y_i^2$$

3.15 设样本 $X=(X_1,\cdots,X_m)$ 和样本 $Y=(Y_1,\cdots,Y_n)$ 互相独立，它们分别来自正态总体 $N(\mu_1,\sigma_1^2)$ 和 $N(\mu_2,\sigma_2^2)$，其中 $-\infty<\mu_1<\infty$，$-\infty<\mu_2<\infty$，$\sigma_1^2>0$，$\sigma_2^2>0$.

(1) 试证明：在 $\sigma_1^2=\sigma_2^2=\sigma^2$ 时，

$$V = \frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}{\sum_{i=1}^{m}(X_i-\overline{X})^2+\sum_{i=1}^{n}(Y_i-\overline{Y})^2} \sim Be\left(\frac{n-1}{2},\frac{m-1}{2}\right)$$

(2) 试证明：检验问题 $H_0: \sigma_1^2=\sigma_2^2$ 对 $H_1: \sigma_1^2\neq\sigma_2^2$ 的 UMPUT 的拒绝域为 $\{x,y: v\leqslant c_1$ 或 $v\geqslant c_2\}$，其中 c_1 和 c_2 由以下两式确定

$$\int_{c_1}^{c_2} Be\left(x\left|\frac{n-1}{2},\frac{m-1}{2}\right.\right)\mathrm{d}x = 1-\alpha$$

$$\int_{c_1}^{c_2}\left(x-\frac{n-1}{m+n-2}\right)\cdot Be\left(x\left|\frac{n-1}{2},\frac{m-1}{2}\right.\right)\mathrm{d}x = 0$$

(3) 试证明：该 UMPUT 的拒绝域可等价地写为 $\{x,y: F\leqslant c_1$ 或 $F\geqslant c_2\}$，其中 c_1 和 c_2 由以下两式确定

$$\int_{c_1}^{c_2} F(x|n-1,m-1)\mathrm{d}x = 1-\alpha$$

$$\int_{c_1}^{c_2}\left(\frac{x-1}{(n-1)x+(m-1)}\right)\cdot F(x|n-1,m-1)\mathrm{d}x = 0$$

3.16 设样本 $X=(X_1,\cdots,X_n)$ 来自双参数指数分布总体，其密度函数为

$$p(x;\mu,\sigma) = \begin{cases} \frac{1}{\sigma}\cdot\exp\left\{-\frac{x-\mu}{\sigma}\right\}, & x\geqslant\mu \\ 0, & x<\mu \end{cases}$$

(1) 试求检验问题 $H_0: \sigma=1$ 对 $H_1: \sigma\neq1$ 的 UMPUT；

(2) 试求检验问题 $H_0: \mu=0$ 对 $H_1: \mu\neq0$ 的 UMPUT.

3.17 (定理 3.18 的补遗) 试证明：

$$I_{ij}(\theta) = \mathrm{E}_\theta\left[\left(\frac{\partial \ln f(X;\theta)}{\partial\theta_i}\right)\cdot\left(\frac{\partial \ln f(X;\theta)}{\partial\theta_j}\right)\right]$$

$$= -E_\theta\left[\frac{\partial^2 \ln f(X;\theta)}{\partial\theta_i \partial\theta_j}\right]$$

3.18　设$(X_1,Y_1),\cdots,(X_n,Y_n)$为来自二元正态分布$N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$的样本，其中$\rho$是相关系数．试求检验问题$H_0:\rho=0$对$H_1:\rho\neq 0$的似然比检验．

3.19　试证明下述结论：

(1)接例 3.16，

$$2\ln\Lambda = 2\sum_{i=1}^{r} n_i \cdot \ln\frac{n_i}{np_{0i}}$$

$$= \sum_{i=1}^{r}\frac{(n_i - np_{0i})^2}{np_{0i}} + \text{依概率收敛于 0 的量}$$

(2)接§3.6.4，

$$2\ln\Lambda = 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \cdot \ln\frac{n \cdot n_{ij}}{n_{i\cdot} \cdot n_{\cdot j}}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - n\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}} + \text{依概率收敛于 0 的量}$$

3.20　设有X,Y二个离散型随机变量．X表示人的性别，“$X=1$”表示男性；“$X=2$”表示女性；Y表示是否色盲，“$Y=1$”表示正常；“$Y=2$”表示色盲．令$p_{ij}=P(X=i,Y=j),i,j=1,2$．这时$\theta=(p_{11},p_{12},p_{21})$，其中$0\leqslant p_{11},p_{12},p_{21}\leqslant 1$，且$p_{11}+p_{12}+p_{21}\leqslant 1$．所以参数空间$\boldsymbol{\Theta}$为$\mathbf{R}^3$中的一个单纯形．

有 1000 人按性别与色盲分类如下

	正常	色盲
男	442	38
女	514	6

按遗传学模型，数据应有下列相对应的概率，

	正常	色盲
男	$\frac{p}{2}$	$\frac{1-p}{2}$
女	$\frac{p^2}{2}+p\cdot(1-p)$	$\frac{(1-p)^2}{2}$

其中 $0\leqslant p\leqslant 1$. 问数据与模型是否相符?

3.21 设有 X,Y,Z 三个离散型随机变量,与此相联系的有一个三维列联表 $r\times c\times t$. 作 n 次观测,在 (i,j,k) 格的观测频数为 n_{ijk}, $i=1,\cdots,r$; $j=1,\cdots,c$, $k=1,\cdots,t$. 观测值落入 (i,j,k) 格的概率为 p_{ijk}.

(1)试证明:在给定 Z 之后,X 和 Y 条件相互独立的时候,p_{ijk} 的 MLE 为

$$\hat{p}_{ijk}=\frac{n_{i.k}\cdot n_{.jk}}{n\cdot n_{..k}}$$

(2) $A=\{(p_{i.k},p_{.jk},p_{..k}): i=1,\cdots,r; j=1,\cdots,c; k=1,\cdots,t;$

$$p_{i.k}\geqslant 0, p_{.jk}\geqslant 0, \sum_{i=1}^{r}p_{i.k}=\sum_{j=1}^{c}p_{.jk}=p_{..k}, \sum_{k=1}^{t}p_{..k}=1\}$$

A 中独立参数为什么有 $t(r+c-1)-1$ 个?

3.22 接 § 3.6.5,试解下列检验问题:

(1)原假设 H_0:给定被告的肤色后被害人的肤色与死刑判决相互条件独立;

(2)原假设 H_0:给定是否死刑判决后被告的肤色与被害人的肤色相互条件独立.

3.23 下表是某大学秋季招生情况的数据. 令 $A=$是否录取,$B=$系别,$C=$性别.

(1)检验:A 和 C 相互独立性. 该大学秋季招生有没有性别歧视?如果有,那种性别的录取率高?

(2)检验:给定 B 之后,A 和 C 的条件独立性. 该大学秋季招生有没有性别歧视?

系别(B)	性别(C)	是否录取(A) 是	否
B_1	男	353	207
	女	17	8
B_2	男	120	205
	女	202	391
B_3	男	138	279
	女	131	244
B_4	男	53	138
	女	94	299
B_5	男	22	351
	女	24	317

3.24　试证明：

(1)以 $\phi(X_1,\cdots,X_m)$为核的 U 统计量的方差不可能比核 $\phi(X_1,\cdots,X_m)$的方差大；

(2)若 U 统计量的核 $\phi(X_1,\cdots,X_m)$的方差有限，则 $\phi_c(X_1,\cdots,X_m)$的方差 σ_c^2 也有限，其中 $\phi_c(X_1,\cdots,X_m)$如(3.73)式所示.

3.25　接例 3.17，在原假设成立时，试证明：

$$\phi_2(x_1,x_2)=\begin{cases}[F(x_1-x_2)+\overline{F}(x_1+x_2)]/3, & x_1>x_2\\ [F(x_2-x_1)+\overline{F}(x_1+x_2)]/3, & x_2>x_1\end{cases}$$

$$\sigma_2^2=\frac{7}{1296}$$

$$\phi_3(x_1,x_2,x_3)=\phi(x_1,x_2,x_3)$$

$$\sigma_3^2=\frac{1}{48}$$

3.26　接例 3.18，在原假设成立时，试证明：

$$\phi_1(x_1)=F(x_1),\qquad \sigma_1^2=\frac{1}{12}$$

$$\phi_2(x_1,x_2)=\phi(x_1,x_2), \qquad \sigma_2^2=\frac{1}{4}$$

$$\mathrm{Var}[U(X_1,\cdots,X_n)]=\frac{2n-1}{6n(n-1)}$$

3.27　接例 3.19,在原假设成立时,试证明:

$$\phi_{10}(x_1)=G(x_1),\phi_{01}(y_1)=\overline{F}(y_1),\phi_{11}(x_1,y_1)=\phi(x_1,y_1)$$

$$\sigma_{10}^2=\sigma_{01}^2=\frac{1}{12},\sigma_{11}^2=\frac{1}{4}.$$

$$\mathrm{Var}\left[U(X_1,\cdots,X_{n_1};Y_1,\cdots,Y_{n_2})\right]=\frac{n_1+n_2+1}{12n_1n_2}$$

3.28　(Kendall τ 检验)设$(X_1,Y_1),\cdots,(X_n,Y_n)$是来自连续总体$(X,Y)$的样本.欲检验假设:原假设 H_0:X 和 Y 相互独立对备择假设 H_1:X 和 Y 正相关. 1938 年 Kendall 构造了以 $\phi((X_1,Y_1)(X_2,Y_2))$为核的 U 统计量,其中 $\phi((X_1,Y_1)(X_2,Y_2))=1$ 或 0,在$(X_1-X_2)(Y_1-Y_2)>0$ 或其它.

(1)在原假设成立时,试证明:$\mathrm{E}(U)=0.5$;

(2)这里将正相关理解为

$$P(Y_1-Y_2>0|X_1-X_2>0)>0.5$$

和
$$P(Y_1-Y_2<0|X_1-X_2<0)>0.5$$

在原假设不成立时,试证明:$\mathrm{E}(U)>0.5$;

(3)试在大样本场合下,给出检验的拒绝域;

(4)上述检验问题可认为是"有方向的"的检验问题,而下述检验问题,原假设 H_0:X 和 Y 相互独立,对备择假设 H_1:X 和 Y 不相互独立,可认为是"无方向的"的检验问题.令

$$\theta=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[F(x,y)-F_x(x)\cdot F_y(y)]^2\mathrm{d}F(x,y)$$

基于样本构造参数 θ 的 U 统计量(假定样本容量足够大),并基于该 U 统计量给出这个"无方向的"的检验问题的解,其中 $F(x,y)$是(X,Y)的联合分布函数,$F_x(x)$和 $F_y(y)$分别是 X 和 Y 的边缘分布函数.

第四章 区间估计

§4.1 基本概念

§4.1.1 区间估计

在第二章中，给出了参数 θ 的估计 $\hat{\theta}=\hat{\theta}(X)$后，我们用估计的均方误差 $\mathrm{E}_\theta[\hat{\theta}(X)-\theta]^2$ 表示估计的精度. 特别地，在 $\hat{\theta}$ 是 θ 的无偏估计时，均方误差就是 $\hat{\theta}$ 的方差. 无偏估计的精度可用它的方差表示. 本章将给出表示精度的另一种方法，那就是给出一个随机区间$[\hat{\theta}_L(X),\hat{\theta}_U(X)]$，例如$[\hat{\theta}(X)-\mathrm{d}(X),\hat{\theta}(X)+\mathrm{d}(X)]$，使得区间$[\hat{\theta}_L(X),\hat{\theta}_U(X)]$包含 θ 的可能性相当大. 用这种方法表示估计的精度既简单又直观. 事实上，区间$[\hat{\theta}_L(X),\hat{\theta}_U(X)]$也是对参数 θ 的一种估计，称为区间估计.

定义 4.1 设$(\mathscr{X},\mathscr{B},\mathscr{P})$为一参数统计结构，其中 $\mathscr{P}=\{P_\theta: \theta\in\Theta\subseteq\mathbf{R}\}$. 假设统计量 $\hat{\theta}_L(X)$和 $\hat{\theta}_U(X)$满足条件：$\hat{\theta}_L(X)\leqslant\hat{\theta}_U(X)$，则称$[\hat{\theta}_L(X),\hat{\theta}_U(X)]$为参数 θ 的一个区间估计.

我们用闭区间给出区间估计. 此外，也可用开区间或半开半闭区间给出区间估计. 具体用什么区间，这没有本质上的差别，可由实际问题的性质而定.

§4.1.2 区间估计的可靠度

在参数真值为 θ 时，自然希望随机区间$[\hat{\theta}_L(X),\hat{\theta}_U(X)]$包含 θ 的概率 $P_\theta\{\hat{\theta}_L(X)\leqslant\theta\leqslant\hat{\theta}_U(X)\}$要大. 一般来说，这个概率与 θ 有关. 所以一个好的区间估计应该对所有的 $\theta\in\Theta$，概率 $P_\theta\{\hat{\theta}_L(X)\leqslant\theta\leqslant\hat{\theta}_U(X)\}$都相当大.

定义 4.2 设$[\hat{\theta}_L(X),\hat{\theta}_U(X)]$为参数$\theta$的一个区间估计，则称区间包含$\theta$的概率在参数空间$\Theta$上的下确界

$$\inf_{\theta\in\Theta} P_\theta\{\hat{\theta}_L(X)\leqslant\theta\leqslant\hat{\theta}_U(X)\}$$

为该区间估计的置信系数.

一个区间估计的置信系数愈大，该区间估计作为未知参数的估计就愈可靠. 但是，构造一个置信系数很大的区间估计并不是一件难事. 这就好比将明天中午的气温估计在-10℃与50℃之间. 这个估计可靠度很高，但它的范围太大，很不精确，没什么用处. 所以一个好的区间估计还有一个精确度的要求. 区间估计的精确度的标准不只一个，常用的标准有以下两个.

§4.1.3 区间估计的精确度

最常用的标准是，区间$[\hat{\theta}_L(X),\hat{\theta}_U(X)]$的平均长度$\mathrm{E}_\theta[\hat{\theta}_U(X)-\hat{\theta}_L(X)]$要短，即区间的范围不能太大，这是符合实际的一项要求.

另一个常用的标准是，设参数真值为θ，在$\theta'\neq\theta$时，自然希望区间$[\hat{\theta}_L(X),\hat{\theta}_U(X)]$包含$\theta'$的概率$P_\theta\{\hat{\theta}_L(X)\leqslant\theta'\leqslant\hat{\theta}_U(X)\}$要小. 区间包含非真值的情况愈少出现愈好，这也是符合实际的一项要求.

例 4.1 设样本$X=(X_1,\cdots,X_n)$来自正态总体$N(\mu,\sigma^2)$，其中$-\infty<\mu<\infty,\sigma^2>0$. μ和σ^2的估计分别是样本均值$\overline{X}$和样本方差$S^2=\frac{1}{n-1}\cdot\sum_{i=1}^{n}(X_i-\overline{X})^2$. 考虑到$\mu$是位置参数，我们采用形如$[\overline{X}-k\cdot S,\overline{X}+k\cdot S]$的区间作$\mu$为的区间估计.

设参数真值为μ. 由于

$$P_\mu\{\overline{X}-k\cdot S\leqslant\mu\leqslant\overline{X}+k\cdot S\}$$
$$=P_\mu\left\{\left|\frac{\sqrt{n}\cdot(\overline{X}-\mu)}{S}\right|\leqslant\sqrt{n}\cdot k\right\}$$

其中$\frac{\sqrt{n}\cdot(\overline{X}-\mu)}{S}\sim t(n-1)$，所以区间包含参数真值$\mu$的概率与$\mu$无关. 因而区间的置信系数为$P\{|t|\leqslant\sqrt{n}\cdot k\}$，其中$t\sim t(n-1)$. 显然，$k$愈大，区间的置信系数也愈大，区间就愈可靠.

由于 $\frac{(n-1)\cdot S^2}{\sigma^2}\sim\chi^2(n-1)$，所以区间的平均长度为

$$2\cdot k\cdot \mathrm{E}_{\sigma^2}(S)=\frac{2\sqrt{2}\cdot k}{\sqrt{n-1}}\cdot\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}\cdot\sigma$$

显然，k 愈大，区间也愈长，区间就愈不精确. 下面考察区间包含非真值 μ' ($\mu'\neq\mu$) 的概率. 该概率为

$$P_\mu\{\overline{X}-k\cdot S\leqslant\mu'\leqslant\overline{X}+k\cdot S\}$$
$$=P\left\{\sqrt{n}\cdot\left(\frac{\mu'-\mu}{S}-k\right)\leqslant t\leqslant\sqrt{n}\cdot\left(\frac{\mu'-\mu}{S}+k\right)\right\}$$

其中 $t\sim t(n-1)$. 显然，k 愈大，区间包含非真值的概率也愈大，区间就愈不精确.

由此例可以看到，在样本容量 n 给定后，为了提高可靠度，可增加 k 值，从而放大了区间，降低了精确度. 反过来，为了提高精确度，可减小 k 值，从而缩短了区间，降低了可靠度. 可靠度与精确度相互制约着. 可靠度与精确度这一对矛盾，犹如假设检验中的犯第一类错误的概率与犯第二类错误的概率这一对矛盾，完全与 Neyman 和 Pearson 的假设检验理论的基本思想相类似，Neyman 建议采取某种折衷方案：在使得置信系数达到一定要求的前提下，寻找精确度尽可能高的区间估计，也就是寻找区间平均长度尽可能短，或者区间包含非真值的概率尽可能小的区间估计. 注意，区间平均长度尽可能短与区间包含非真值的概率尽可能小，这两个要求可能同时达到，也可能不同时达到.

§4.1.4　置信水平

在 Neyman 的这种思想指导下，寻找一个好的区间估计，就是对选定的一个较小的数 α ($0<\alpha<1$)，在置信系数不小于 $1-\alpha$，即满足

$$P_\theta\{\hat{\theta}_L(X)\leqslant\theta\leqslant\hat{\theta}_U(X)\}\geqslant 1-\alpha,\quad\forall\,\theta\in\Theta \tag{4.1}$$

的区间估计中，寻找这样的区间估计，使得区间的平均长度 $\mathrm{E}_\theta[\hat{\theta}_U(X)-\hat{\theta}_L(X)]$ 尽可能短，或者使得区间包含非真值的概率 $P_\theta\{\hat{\theta}_L(X)\leqslant\theta'\leqslant\hat{\theta}_U(X)\}$ 尽可能小，其中 $\theta'\neq\theta$.

定义 4.3 满足(4.1)式的区间估计称为置信水平为 $1-\alpha$ 的置信区间，或简称为 $1-\alpha$ 置信区间.

α 是事先选定的. 我们常取 0.1，0.05，0.01 等作为 α 的值. 与假设检验相类似，对给定的置信水平 $1-\alpha$，人们总是力图构造这样的区间估计，使得(4.1)式得到满足，并且至少存在一个 $\theta\in\Theta$，使得

$$P_\theta\{\hat{\theta}_L(X)\leqslant\theta\leqslant\hat{\theta}_U(X)\}=1-\alpha$$

这意味着置信水平被"足量"地使用了，以满足区间的精确度尽可能高的要求. 为此，我们在例 4.1 中取 $k=\frac{t_{1-\alpha/2}}{\sqrt{n}}$，那么

$$\left[\overline{X}-\frac{S}{\sqrt{n}}\cdot t_{1-\alpha/2}(n-1),\overline{X}+\frac{S}{\sqrt{n}}\cdot t_{1-\alpha/2}(n-1)\right]$$

就是正态均值 μ 的置信水平 $1-\alpha$ 被"足量"地使用的置信区间.

譬如，从正态总体 $N(\mu,\sigma^2)$ 抽取容量为 10 的一个样本. 设样本观察值为

78.1　72.4　76.2　74.3　77.4　78.4　76.0　75.5　76.7　77.3

由此算得 $\overline{X}=76.23$，$S=1.82$. 若要求 μ 的 95%置信区间，这就是说 $1-\alpha=0.95$，$\frac{\alpha}{2}=0.025$. 由 t 分布表查得 $t_{0.975}(9)=2.26$. 代入

$$\left[\overline{X}-\frac{S}{\sqrt{10}}\cdot t_{0.975}(9),\overline{X}+\frac{S}{\sqrt{10}}\cdot t_{0.975}(9)\right] \tag{4.2}$$

可算得

$$\left[76.23-\frac{1.82}{\sqrt{10}}\cdot 2.26,76.23+\frac{1.82}{\sqrt{10}}\cdot 2.26\right]$$

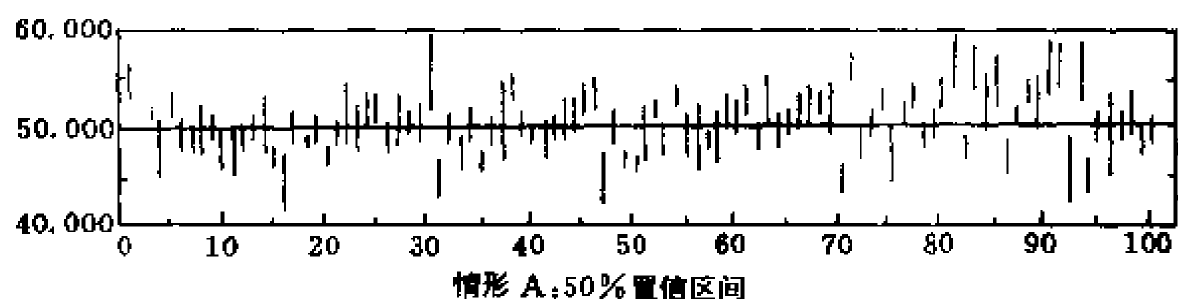
即

$$[76.23-1.30,76.23+1.30]$$

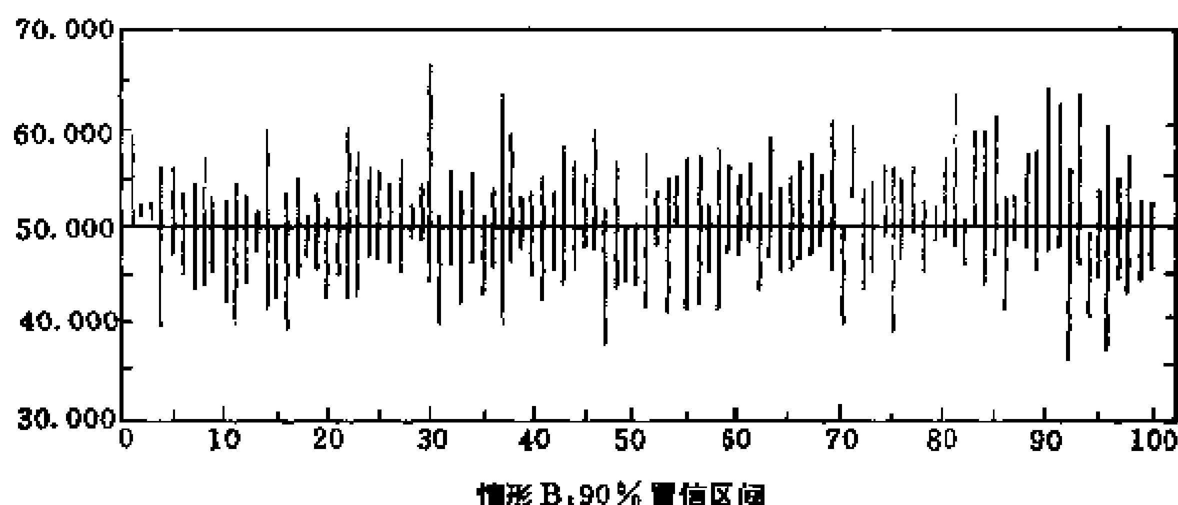
所以 μ 的 95%置信区间为[74.93，77.53].

确切地说，[74.93，77.53]是 μ 的 95%置信区间(4.2)用在一个具体样本时的一次实现. 必须指出的是，虽然 μ 未知，但它是一个常数. 区间[74.93，77.53]内要么含有 μ，要么不含有 μ，两者必居其一，无概率可言. 置信水平为 95%是对(4.2)所表示的置信区间而言的. 它的含意是：如果将(4.2)反复用在一个个具体样本上，在这很多个实现中，大

约有 95%个实现含有 μ,大约有 5%个实现不含有 μ. 图 4.1 是一个随机模拟的结果. 从均值 $\mu=50\ 000$ 和 $\sigma=5\ 000$ 的正态总体中随机取出 100 个容量为 4 的样本. 图上的每一条竖线表示由一个具体样本算得的一个置信区间. 情形 A 表示 50%置信区间,其中大约有 50 个置信区间包含真实参数 μ;情形 B 表示 90%置信区间,其中大约有 90 个置信区间包含真实参数 μ. 情形 A 和 B 分别是置信水平为 50%和 90%的直观解释.



情形 A:50%置信区间



情形 B:90%置信区间

图 4.1　对从均值 $\mu=50\ 000$ 和 $\sigma=5000$ 的正态总体中随机取出 100 个容量为 4 的样本计算得到的置信区间.
情形 A 表示 50%置信区间;情形 B 表示 90%置信区间.

例 4.2　设样本 $X=(X_1,\cdots,X_n)$ 来自正态总体 $N(\mu,\sigma^2)$,其中 $-\infty<\mu<\infty,\sigma^2>0$. 考虑到 σ^2 为尺度参数,我们采用形如 $[aS^2,bS^2]$ 的区间作为 σ^2 的区间估计,其中 $0<a<b,S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$ 为样

本方差.

设参数真值为 σ^2，由于 $\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$，所以若 a 和 b 满足下述条件

$$P_{\sigma^2}\{aS^2\leqslant\sigma^2\leqslant bS^2\}=P_{\sigma^2}\left\{\frac{n-1}{b}\leqslant\frac{(n-1)S^2}{\sigma^2}\leqslant\frac{n-1}{a}\right\}$$

$$=\int_{\frac{n-1}{b}}^{\frac{n-1}{a}}\chi^2(x\,|\,n-1)\mathrm{d}x=1-\alpha \tag{4.3}$$

则区间 $[aS^2,bS^2]$ 就是 σ^2 的 $1-\alpha$ 置信区间，其中 $\chi^2(x\,|\,n-1)$ 表示自由度为 $n-1$ 的 χ^2 分布的密度函数.

该区间的平均长度为

$$L=\frac{(b-a)\sigma^2}{n-1}\cdot\mathrm{E}_{\sigma^2}\left\{\frac{(n-1)S^2}{\sigma^2}\right\}=(b-a)\sigma^2$$

根据 Neyman 的区间估计理论的基本思想，我们需要解下述的条件极值问题：寻找 a 和 b，使得

$$\begin{cases}a\text{ 和 }b\text{ 满足(4.3)式}\\ \min\limits_{a,b}(b-a)\end{cases}$$

由(4.3)式知，b 可以看作是 a 的函数. 记 $b=b(a)$. 令 $f(a)=b(a)-a$. 在(4.3)式的两边对 a 求导数. 由此可以推得，

$$\frac{1}{a^2}\cdot\chi^2\left(\frac{n-1}{a}\,\middle|\,n-1\right)-\frac{1}{b^2}\cdot\chi^2\left(\frac{n-1}{b}\,\middle|\,n-1\right)\cdot b'(a)=0$$

从而由 $f'(a)=b'(a)-1=0$ 推得

$$\frac{1}{a^2}\cdot\chi^2\left(\frac{n-1}{a}\,\middle|\,n-1\right)=\frac{1}{b^2}\cdot\chi^2\left(\frac{n-1}{b}\,\middle|\,n-1\right) \tag{4.4}$$

所以在形如 $[aS^2,bS^2]$ 的 $1-\alpha$ 置信区间中，由(4.3)和(4.4)两式确定的 a 和 b，区间的平均长度达到最短.

在形如 $[aS^2,bS^2]$ 的 $1-\alpha$ 置信区间中，区间包含非真值的概率尽可能小的问题留待后面讨论.

在实用中常取 a 和 b 满足下列条件以致于(4.3)式成立.

$$\int_{\frac{n-1}{a}}^{\infty} \chi^2(x|n-1)\mathrm{d}x = \frac{\alpha}{2}, \quad \int_{-\infty}^{\frac{n-1}{b}} \chi^2(x|n-1)\mathrm{d}x = \frac{\alpha}{2}$$

即取 a 和 b 满足条件：

$$\frac{n-1}{a} = \chi^2_{1-\frac{\alpha}{2}}(n-1), \quad \frac{n-1}{b} = \chi^2_{\frac{\alpha}{2}}(n-1)$$

所以在实用中常取 σ^2 的 $1-\alpha$ 置信区间为

$$\left[\frac{(n-1)\cdot S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}, \frac{(n-1)\cdot S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right]$$

§4.1.5 置信限

在实际问题中，人们有时仅对未知参数的置信下限或置信上限感兴趣. 譬如，电视机的寿命要求愈大愈好，因此人们关心的仅是某种型号的电视机的平均寿命的置信下限，因为它的大小标志着电视机的质量的好坏. 又如，药物的副作用要求愈小愈好，因此人们关心的仅是某种药物的副作用的置信上限，因为它的大小标志着该药物的质量的好坏. 对这些问题去寻求两端都为有界的置信区间就没有必要了.

定义 4.4 设有统计量 $\hat{\theta}_L(X)$. 如果对选定的一个较小的数 $\alpha(0<\alpha<1)$，有

$$P_\theta\{\theta \geqslant \hat{\theta}_L(X)\} \geqslant 1-\alpha, \quad \forall\, \theta \in \Theta \tag{4.5}$$

则 $\hat{\theta}_L(X)$ 称为 θ 的（置信水平为）$1-\alpha$（单侧）置信下限. $\inf\limits_{\theta\in\Theta} P_\theta\{\theta \geqslant \hat{\theta}_L(X)\}$ 称为置信下限的置信系数. 类似地，设统计量 $\hat{\theta}_U(X)$ 满足下述条件，

$$P_\theta\{\theta \leqslant \hat{\theta}_U(X)\} \geqslant 1-\alpha, \quad \forall\, \theta \in \Theta \tag{4.6}$$

则 $\hat{\theta}_U(X)$ 称为 θ 的（置信水平为）$1-\alpha$（单侧）置信上限. $\inf\limits_{\theta\in\Theta} P_\theta\{\theta \leqslant \hat{\theta}_U(X)\}$ 称为置信上限的置信系数.

显然，对置信下限 $\hat{\theta}_L(X)$ 来说，$\mathrm{E}_\theta[\hat{\theta}_L(X)]$ 愈大，置信下限的精度愈高；而对置信上限 $\hat{\theta}_U(X)$ 来说，$\mathrm{E}_\theta[\hat{\theta}_U(X)]$ 愈小，置信上限的精度愈高. 此外，若参数真值为 θ，则对置信下限 $\hat{\theta}_L(X)$ 来说，在 $\theta'<\theta$ 时，θ' 大于或

等于 $\hat{\theta}_L(X)$ 的概率 $P_\theta\{\theta'\geqslant\hat{\theta}_L(X)\}$ 愈小，置信下限的精度愈高. 同样地，对置信上限 $\hat{\theta}_U(X)$ 来说，在 $\theta'>\theta$ 时，θ' 小于或等于 $\hat{\theta}_U(x)$ 的概率 $P_\theta\{\theta'\leqslant\hat{\theta}_U(X)\}$ 愈小，置信上限的精度愈高.

根据 Neyman 的区间估计的基本思想，寻找一个好的置信下限(或置信上限)，就是对选定的一个较小的数 $\alpha(0<\alpha<1)$，在置信水平为 $1-\alpha$ 的置信下限(或置信上限)中，寻找精度尽可能高的置信下限(或置信上限).

§4.1.6 置信域

置信区间的概念可以推广到多个参数的情况.

定义 4.5 设 $(\mathscr{X},\mathscr{B},\mathscr{P})$ 为一参数统计结构，其中 $\mathscr{P}=\{P_\theta:\theta\in\Theta\subseteq\mathbf{R}^k\}$，$\theta=(\theta_1,\cdots,\theta_k)$. 假设 $S(X)$ 满足下述条件

(1) 对任意一个样本观察值 x，$S(x)$ 是 Θ 的一个子集合；

(2) 对给定的 $\alpha(0<\alpha<1)$，$P_\theta\{\theta\in S(X)\}\geqslant1-\alpha$, $\forall\,\theta\in\Theta$.

则称 $S(X)$ 是 θ 的(置信水平为)$1-\alpha$ 置信域(或置信集)，而概率 $P_\theta\{\theta\in S(X)\}$ 在参数空间 Θ 上的下确界称为置信域的置信系数.

置信域的形状可能是多种多样的，但常用的是一些有规则的几何图形，如长方体、球、椭球等. 在置信域是长方体，且长方体的各个面与坐标平面平行时，此置信域又称为联合区间估计.

本章着重讨论单个参数的置信区间和置信限的问题，至于多个参数的置信域的问题，可参阅有关多元统计分析的教材.

§4.2 构造置信区间(置信限)的方法

§4.2.1 枢轴量法

设 $(\mathscr{X},\mathscr{B},\mathscr{P})$ 为一参数统计结构，其中 $\mathscr{P}=\{P_\theta:\theta\in\Theta\subseteq\mathbf{R}^k\}$. 可以按下列三个步骤构造 $\eta=g(\theta)$ 的置信区间(限).

(1) 构造样本 X 和参数 η 的一个函数 $G=G(X,\eta)$，要求 G 的分布

与 η 无关. 具有这种性质的函数称为枢轴量.

(2) 对给定的 $\alpha(0<\alpha<1)$, 选取两个常数 c 和 $d(c<d)$, 使得

$$P_\theta\{c \leqslant G(X,\eta) \leqslant d\} \geqslant 1-\alpha, \quad \forall\ \theta \in \Theta \tag{4.7}$$

(3) 如果不等式 $c \leqslant G(X,\eta) \leqslant d$ 可等价地变换为 $\hat{\eta}_L(X) \leqslant \eta \leqslant \hat{\eta}_U(X)$, 那么

$$P_\theta\{\hat{\eta}_L(X) \leqslant \eta \leqslant \hat{\eta}_U(X)\} \geqslant 1-\alpha, \quad \forall\ \theta \in \Theta \tag{4.8}$$

则 $[\hat{\eta}_L(X), \hat{\eta}_U(X)]$ 为 η 的一个置信水平为 $1-\alpha$ 的置信区间. 在 $G(X,\eta)$ 是 η 的连续、严格单调函数时, 这两个不等式的等价变换总是可以做到的.

类似地, 选取常数 c(或 d), 使得 $P_\theta\{c \leqslant G(X,\eta)\} \geqslant 1-\alpha$(或 $P_\theta\{d \geqslant G(X,\eta)\} \geqslant 1-\alpha$), 可以构造出 η 的置信限.

例 4.3　设样本 $X=(X_1,\cdots,X_n)$ 来自均匀分布总体 $U(0,\theta)$, 其中 $\theta>0$. 对给定的 $\alpha(0<\alpha<1)$, 欲构造 θ 的置信水平为 $1-\alpha$ 的置信区间.

解: 由于最大次序统计量 $X_{(n)}=\max\{X_1,\cdots,X_n\}$ 是 θ 的 MLE, 也是 θ 的充分统计量, 所以我们基于最大次序统计量 $X_{(n)}$ 构造 θ 的区间估计的想法是可取的. 因为 $\dfrac{X_{(n)}}{\theta}$ 的密度函数为 $n \cdot y^{n-1}(0 \leqslant y \leqslant 1)$, 故取枢轴量 $G(X,\theta)=\dfrac{X_{(n)}}{\theta}$. 对给定的 $\alpha(0<\alpha<1)$, 只要 a 和 $b(a<b)$ 满足条件: $b^n-a^n=1-\alpha$, 就有 $P_\theta\left(a \leqslant \dfrac{X_{(n)}}{\theta} \leqslant b\right)=1-\alpha$. 由不等式等价变形可以得到 θ 的置信水平为 $1-\alpha$ 的置信区间 $\left[\dfrac{X_{(n)}}{b}, \dfrac{X_{(n)}}{a}\right]$. 这里我们仅考虑区间的平均长度愈短愈好的精度要求. 则取 $b=1, a=\sqrt[n]{\alpha}$. 从而在形如 $\left[\dfrac{X_{(n)}}{b}, \dfrac{X_{(n)}}{a}\right]$ 的置信水平为 $1-\alpha$ 的置信区间中, 区间 $\left[X_{(n)}, \dfrac{X_{(n)}}{\sqrt[n]{\alpha}}\right]$ 的平均长度达到最短(为什么? 见习题 4.1).

设 8.3, 5.0, 1.7, 2.9, 4.3, 1.8, 4.6, 5.8, 1.6, 6.6 是取自均匀分布 $U(0,\theta)$ 的十个观察值. 给定 $\alpha=0.05$, 则 θ 的置信水平为 95% 的置信区间为 [8.3, 11.2].

在例 4.1 中,我们讨论了单个正态总体均值 μ 的区间估计问题. 如果取 $t=\sqrt{n}\cdot\frac{\bar{X}-\mu}{S}\sim t(n-1)$ 作为枢轴量,构造 μ 的置信区间(限)就比较容易了. 在例 4.2 中,我们讨论了单个正态总体方差 σ^2 的区间估计问题. 如果取 $\chi^2=\frac{1}{\sigma^2}\cdot\sum_{i=1}^{n}(X_i-\bar{X})^2\sim\chi^2(n-1)$ 作为枢轴量,构造 σ^2 的置信区间(限)也就比较容易了.

例 4.4 设样本 $X=(X_1,\cdots,X_m)$ 和 $Y=(Y_1,\cdots,Y_n)$ 互相独立. 它们分别来自正态总体 $N(\mu_1,\sigma^2)$ 和 $N(\mu_2,\sigma^2)$,其中 $-\infty<\mu_1<\infty$, $-\infty<\mu_2<\infty$, $\sigma^2>0$. 这是方差相等的两个正态总体. 现讨论它们的均值差 $\delta=\mu_1-\mu_2$ 的区间估计问题.

取

$$t=\sqrt{\frac{mn(m+n-2)}{m+n}}\cdot\frac{\bar{X}-\bar{Y}-\delta}{\sqrt{\sum_{i=1}^{m}(X_i-\bar{X})^2+\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}\sim t(m+n-2)$$

作为枢轴量,不难得到 δ 的置信水平为 $1-\alpha$ 的置信区间或置信限.

例 4.5 设样本 $X=(X_1,\cdots,X_m)$ 和 $Y=(Y_1,\cdots,Y_n)$ 互相独立. 它们分别来自正态总体 $N(\mu_1,\sigma_1^2)$ 和 $N(\mu_2,\sigma_2^2)$,其中 $-\infty<\mu_1<\infty$, $-\infty<\mu_2<\infty$, $\sigma_1^2>0$, $\sigma_2^2>0$. 现讨论这两个正态总体的方差之比 $\rho=\frac{\sigma_1^2}{\sigma_2^2}$ 的区间估计问题.

取

$$F=\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{\sum_{i=1}^{m}(X_i-\bar{X})^2}\cdot\frac{m-1}{n-1}\cdot\rho\sim F(n-1,m-1)$$

作为枢轴量,不难得到 ρ 的置信水平为 $1-\alpha$ 的置信区间或置信限.

在样本容量充分大时,有时可用渐近分布来构造近似的置信区间. 我们用下面的例子来说明这个方法.

例 4.6 设样本 $X=(X_1,\cdots,X_n)$ 来自二点分布总体 $b(1,p)$，其中 $0\leqslant p\leqslant 1$. 欲求 p 的置信区间.

解：p 的 MLE 为 $\hat{p}=\overline{X}$. 由中心极限定理知，在 $n\to\infty$ 时，

$$\sqrt{n}\cdot\frac{\hat{p}-p}{\sqrt{p(1-p)}}\xrightarrow{L}N(0,1)$$

所以在样本容量 n 充分大时，

$$P\left\{-\lambda\leqslant\sqrt{n}\cdot\frac{\hat{p}-p}{\sqrt{p(1-p)}}\leqslant\lambda\right\}\approx 1-\alpha \tag{4.9}$$

其中 $\lambda=U_{1-\alpha/2}$. 由此得 p 的置信水平近似为 $1-\alpha$ 的置信区间 $[\hat{p}_L,\hat{p}_U]$，其中

$$\hat{p}_L=\frac{n}{n+\lambda^2}\cdot\left[\hat{p}+\frac{\lambda^2}{2n}-\lambda\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{\lambda^2}{4n^2}}\right]$$

$$\hat{p}_U=\frac{n}{n+\lambda^2}\cdot\left[\hat{p}+\frac{\lambda^2}{2n}+\lambda\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{\lambda^2}{4n^2}}\right]$$

在实用中，由 $\hat{p}\xrightarrow{P}p\,(n\to\infty)$ 知

$$\sqrt{n}\cdot\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})}}\xrightarrow{L}N(0,1),\quad n\to\infty$$

从而在样本容量 n 充分大时，我们有

$$P\left\{-\lambda\leqslant\sqrt{n}\cdot\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})}}\leqslant\lambda\right\}\approx 1-\alpha \tag{4.10}$$

由此得 p 的置信水平近似为 $1-\alpha$ 的置信区间为

$$\left[\hat{p}-\lambda\cdot\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}},\hat{p}+\lambda\cdot\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right] \tag{4.11}$$

在实际问题中常用区间估计(4.11)，它较为简单. 这个区间估计经过了两次近似. (4.9)式是第一次近似. 从(4.9)式到(4.10)式，把分母中的 p 换为 $\hat{p}$，这是第二次近似. 所以若用区间估计(4.11)，样本容量 n 必须较大，否则置信水平与 $1-\alpha$ 有较大的差距. 一般要求 n 不小于 30.

区间估计(4.11)可用来确定样本容量. 譬如，某国(或某地区)要了

解公众对某项政策的支持率 p. 如果要求估计值 $\hat{p}$ 与真值 p 之间的最大误差不超过 3%，且置信系数为 95%，则应随机地调查多少居民？由于居民很多，所以可认为总体是二点分布 $b(1,p)$. 取 $\lambda=U_{0.975}=1.96$，则由(4.11)式知

$$P\left\{\hat{p}-1.96\cdot\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\leqslant p\leqslant\hat{p}+1.96\cdot\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right\}\approx 0.95$$

从而有

$$P\left\{|\hat{p}-p|\leqslant 1.96\cdot\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right\}\approx 0.95$$

由于 $0\leqslant\hat{p}\leqslant 1$，所以 $\hat{p}(1-\hat{p})\leqslant 0.25$. 则为了使得 $P\{|\hat{p}-p|\leqslant 3\%\}\approx 0.95$，只须样本容量 n 满足下述条件：$1.96\cdot\frac{\sqrt{0.25}}{\sqrt{n}}\leqslant 3\%$ 即可. 这也就是说 $n\geqslant\left[1.96\cdot\frac{\sqrt{0.25}}{3\%}\right]^2\approx 1067$. 所以即使一个国家的居民上亿，在做关于支持率的民意调查时，只要随机访问 1067 个居民，就可使得支持率的估计值与真值之间的最大误差不超过 3%，且置信系数近似为 95%.

§4.2.2　基于连续随机变量构造置信区间

一般来说，欲构造 θ 的置信水平为 $1-\alpha$ 的置信区间，我们首先考虑 θ 的 MLE，或 θ 的充分统计量. 然后据此得到了一个统计量 $T=T(X)$，并基于 $T(X)$ 寻找枢轴量，然后构造 θ 的置信区间，或大样本时构造 θ 的近似的置信区间.

在统计量 $T(X)$ 是连续随机变量时，寻找枢轴量不是一件难事. 设 $T(X)$ 的分布函数为 $G(t,\theta)=P_\theta\{T(X)\leqslant t\}$，那么 $G(T(X),\theta)$ 服从区间 $(0,1)$ 上的均匀分布. 这时 $G(T(X),\theta)$ 可被取为枢轴量. 具体做法看下面的例子.

例 4.7　设 $X=(X_1,\cdots,X_n)$ 是来自密度函数为

$$p(x;\theta)=\frac{\theta}{x^2},\quad 0<\theta\leqslant x<\infty$$

总体的样本. 试求 θ 的置信水平为 $1-\alpha$ 的置信区间.

解:最小次序统计量 $X_{(1)}=\min\{X_1,\cdots,X_n\}$ 是 θ 的 MLE,也是 θ 的充分统计量,所以我们将基于最小次序统计量 $X_{(1)}$ 构造 θ 的区间估计.

$T=X_{(1)}$ 的密度函数和分布函数分别为

$$g(t,\theta)=\frac{n\cdot\theta^n}{t^{n+1}},\quad 0<\theta\leqslant t<\infty$$

和

$$G(t,\theta)=1-\frac{\theta^n}{t^n},\quad 0<\theta\leqslant t<\infty$$

则枢轴量可取为

$$G(X_{(1)},\theta)=1-\frac{\theta^n}{X_{(1)}^n}\sim U(0,1)$$

选取两个常数 c 和 $d(0<c<d<1)$,使得 $d-c=1-\alpha$,则有

$$P_\theta\left\{c\leqslant 1-\frac{\theta^n}{X_{(1)}^n}\leqslant d\right\}=1-\alpha$$

从而有 θ 的置信水平为 $1-\alpha$ 的置信区间 $\left[\sqrt[n]{1-d}\cdot X_{(1)},\sqrt[n]{1-c}\cdot X_{(1)}\right]$. 考虑到区间的平均长度愈短愈好的精度要求,如何选取 c 和 d 的问题见习题 4.2.

在 $T(X)$ 是连续随机变量时,$G(T(X),\theta)$ 可被取为枢轴量,其中 $G(t,\theta)=P_\theta\{T(X)\leqslant t\}$ 是 $T(X)$ 的分布函数. 用枢轴量构造 θ 的置信区间的第三个步骤中,不等式 $c\leqslant G(T(X),\theta)\leqslant d$ 可等价地变换成 $\hat{\theta}_L(T(X))\leqslant\theta\leqslant\hat{\theta}_U(T(X))$ 的关键之处在于,$T(X)$ 的分布函数 $G(t,\theta)$ 是 θ 的单调函数. 这在 $T=T(X)$ 的概率密度族关于 $T(X)$ 具有单调似然比(MLR)时,总是可以做到的. 这时 $G(t,\theta)$ 是 θ 的一个非增(或非降),甚至严格单调的函数(见定理 3.7 和定理 3.8(2)).

§4.2.3 基于离散随机变量构造置信区间

如果统计量 $T(X)$ 是离散随机变量,那么 $G(T(X),\theta)$ 就不服从区间(0,1)上的均匀分布,其中 $G(t,\theta)=P_\theta\{T(X)\leqslant t\}$ 是 $T(X)$ 的分布函数. 这时关于 $G(T(X),\theta)$ 有下面的结论,据此可以构造出 θ 的置信区间.

引理 4.1 设 $F(x)$是随机变量 X 的分布函数.若 $0\leqslant y\leqslant 1$,则有

$$P\{F(X)\leqslant y\}\leqslant y\leqslant P\{F(X-0)<y\}$$

证明:首先分下面两种情况证明左边的不等式.

情况 1:设集合$\{x:F(x)=y\}$非空,并且该集合的上确界可以达到,即

$$x_0=\sup\{x:F(x)=y\}\in\{x:F(x)=y\}$$

则有

$$P\{F(X)\leqslant y\}=P\{X\leqslant x_0\}=F(x_0)=y$$

左边的不等式成立.

情况 2:设集合$\{x:F(x)=y\}$是空集,或者该集合虽然非空,但它的上确界不能达到.则函数 $F(x)$存在这样的一点 x,使得 $F(x-0)\leqslant y<F(x)$.则有

$$P\{F(X)\leqslant y\}=P\{F(X)<F(x)\}=P\{X<x\}=F(x-0)\leqslant y$$

这就证明了左边的不等式.

现在证明右边的不等式.令 $Z=-X$,并设 Z 的分布函数为 $G(z)$.根据已证明了的结果,有 $P\{G(Z)\leqslant 1-y\}\leqslant 1-y$.由于

$$G(z)=P\{Z\leqslant z\}=P\{X\geqslant -z\}=1-F(-z-0)$$

所以

$$1-y\geqslant P\{F(-Z-0)\geqslant y\}=1-P\{F(X-0)<y\}$$

从而有 $P\{F(X-0)<y\}\geqslant y$.右边的不等式得证.引理证毕.

如果 $F(x)$是连续函数,则据引理 4.1,有 $P\{F(X)\leqslant y\}=y(0\leqslant y\leqslant 1)$.这说明 $F(X)\sim U(0,1)$.这是关于连续随机变量的著名结论.所以引理 4.1 是把这个结论推广到任意分布函数上去得到的结果.

如果 $T(X)$是离散随机变量,其分布函数为 $G(t,\theta)=P_\theta\{T(x)\leqslant t\}$,则由引理 4.1 知,

$$P_\theta\{G(T(X),\theta)\leqslant y\}\leqslant y\leqslant P_\theta\{G(T(X)-0,\theta)<y\}$$

对任意的 $0\leqslant y\leqslant 1$ 和一切的 $\theta\in\Theta$ 皆成立.据此可以构造出 θ 的置信区间.

设$(\mathscr{X},\mathscr{B},\mathscr{P})$为一单参数统计结构,其中 $\mathscr{P}=\{P_\theta:\theta\in\Theta\subseteq\mathbf{R}\}$.设 $X=(X_1,\cdots,X_n)$是来自该单参数统计结构的一个样本.对给定的

$\alpha(0<\alpha<1)$，欲构造 θ 的置信水平为 $1-\alpha$ 的置信区间. 首先，我们考虑 θ 的 MLE，或 θ 的充分统计量. 然后据此得到了一个统计量 $T=T(X)$，并基于 $T(X)$ 构造 θ 的置信区间. 设 $T=T(X)$ 的分布函数为 $G(t,\theta)$.

定理 4.2 如果 $G(t,\theta)$ 是 θ 的严格减函数，则对给定的 $\alpha(0<\alpha<1)$，

$$\hat{\theta}_L=\sup_{\theta\in\Theta}\{\theta: G(T-0,\theta)\geqslant 1-\alpha\} \tag{4.12}$$

$$\hat{\theta}_U=\inf_{\theta\in\Theta}\{\theta: G(T,\theta)\leqslant \alpha\} \tag{4.13}$$

分别是 θ 的置信水平为 $1-\alpha$ 的置信下限和 $1-\alpha$ 的置信上限.

证明：由于 $G(t,\theta)$ 是 θ 的严格减函数，所以在 $G(T-0,\theta)<1-\alpha$ 时，必然有 $\theta\geqslant\hat{\theta}_L$. 则据引理4.1知

$$P_\theta\{\theta\geqslant\hat{\theta}_L\}\geqslant P_\theta\{G(T-0,\theta)<1-\alpha\}\geqslant 1-\alpha$$

(4.12)得证. 同样地，在 $G(T,\theta)>\alpha$ 时，必然有 $\theta\leqslant\hat{\theta}_U$，从而据引理 4.1 知，

$$P_\theta\{\theta\leqslant\hat{\theta}_U\}\geqslant P_\theta\{G(T,\theta)>\alpha\}=1-P_\theta\{G(T,\theta)\leqslant\alpha\}\geqslant 1-\alpha$$

(4.13)得证. 证毕.

利用定理 4.2，即得下面的定理.

定理 4.3 如果 $G(t,\theta)$ 是 θ 的严格减函数，则对给定的 $\alpha(0<\alpha<1)$，θ 的置信水平为 $1-\alpha$ 的置信区间为 $[\hat{\theta}_L,\hat{\theta}_U]$，其中

$$\hat{\theta}_L=\sup_{\theta\in\Theta}\{\theta: G(T-0,\theta)\geqslant 1-\alpha_1\},\hat{\theta}_U=\inf_{\theta\in\Theta}\{\theta: G(T,\theta)\leqslant \alpha_2\}$$

这里的 $\alpha_1+\alpha_2=\alpha$，且 $0\leqslant\alpha_1,\alpha_2\leqslant 1$.

常根据区间的平均长度愈短愈好的精度要求，选取 α_1 和 α_2. 在根据这个要求选取 α_1 和 α_2 有困难，或较为复杂时，可取 $\alpha_1=\alpha_2=\dfrac{\alpha}{2}$.

如果 $G(t,\theta)$ 是 θ 的严格增函数，那么 $G(t,\theta)$ 是 $-\theta$ 的严格减函数，由定理 4.2 和定理 4.3 可以得到 $-\theta$ 的置信区间(限)，从而有下述定理.

定理 4.4 若 $G(t,\theta)$ 是 θ 的严格增函数，

(1)则对给定的 $\alpha(0<\alpha<1)$，

$$\hat{\theta}_L=\inf_{\theta\in\Theta}\{\theta: G(T,\theta)\leqslant \alpha\}$$

$$\hat{\theta}_U=\sup_{\theta\in\Theta}\{\theta: G(T-0,\theta)\geqslant 1-\alpha\}$$

分别是 θ 的置信水平为 $1-\alpha$ 的置信下限和 $1-\alpha$ 的置信上限；

(2)则对给定的 $\alpha(0<\alpha<1)$，θ 的置信水平为 $1-\alpha$ 的置信区间为 $[\hat{\theta}_L,\hat{\theta}_U]$，其中

$$\hat{\theta}_L=\inf_{\theta\in\Theta}\{\theta: G(T,\theta)\leqslant\alpha_2\},\quad \hat{\theta}_U=\sup_{\theta\in\Theta}\{\theta: G(T-0,\theta)\geqslant 1-\alpha_1\}$$

这里的 $\alpha_1+\alpha_2=\alpha$，且 $0\leqslant\alpha_1,\alpha_2\leqslant 1$.

在 $G(t,\theta)$ 是 θ 的连续严格减函数时，关于 θ 的方程 $G(T-0,\theta)=1-\alpha$ 和 $G(T,\theta)=\alpha$ 的解分别是集合 $\{\theta: G(T-0,\theta)\geqslant 1-\alpha\}$ 的上确界和集合 $\{\theta: G(T,\theta)\leqslant\alpha\}$ 的下确界. 因此上述两个定理可以得到简化.

定理 4.5 如果 $G(t,\theta)$ 是 θ 的连续、严格减函数，那么

(1)关于 θ 的方程 $G(T-0,\theta)=1-\alpha$ 的解是 θ 的置信水平为 $1-\alpha$ 的置信下限；

(2)关于 θ 的方程 $G(T,\theta)=\alpha$ 的解是 θ 的置信水平为 $1-\alpha$ 的置信上限；

(3)θ 的置信水平为 $1-\alpha$ 的置信区间为 $[\hat{\theta}_L,\hat{\theta}_U]$，其中 $\hat{\theta}_L$ 和 $\hat{\theta}_U$ 分别是关于 θ 的方程 $G(T-0,\theta)=1-\alpha_1$ 和 $G(T,\theta)=\alpha_2$ 的解，这里的 $\alpha_1+\alpha_2=\alpha$，且 $0\leqslant\alpha_1,\alpha_2\leqslant 1$.

定理 4.6 如果 $G(t,\theta)$ 是 θ 的连续、严格增函数，那么

(1)关于 θ 的方程 $G(T,\theta)=\alpha$ 的解是 θ 的置信水平为 $1-\alpha$ 的置信下限；

(2)关于 θ 的方程 $G(T-0,\theta)=1-\alpha$ 的解是 θ 的置信水平为 $1-\alpha$ 的置信上限；

(3)θ 的置信水平为 $1-\alpha$ 的置信区间为 $[\hat{\theta}_L,\hat{\theta}_U]$，其中 $\hat{\theta}_L$ 和 $\hat{\theta}_U$ 分别是关于 θ 的方程 $G(T,\theta)=\alpha_1$ 和 $G(T-0,\theta)=1-\alpha_2$ 的解，这里的 $\alpha_1+\alpha_2=\alpha$，且 $0\leqslant\alpha_1,\alpha_2\leqslant 1$.

定理 4.5 和定理 4.6 的证明从略.

例 4.8 设样本 $X=(X_1,\cdots,X_n)$ 来自二点分布总体 $b(1,p)$，其中 $0\leqslant p\leqslant 1$. 欲求 p 的置信区间.

解：在例 4.5 中，我们在样本容量 n 充分大的时候，给出了 p 的置

信水平近似为 $1-\alpha$ 的置信区间. 现在使用定理 4.5 可得到 p 的精确的置信区间.

由于 p 的 MLE 为 $\overline{X}$, 并且 p 的充分统计量为 $T=\sum_{i=1}^{n}X_i$, 所以我们将基于 T 构造 p 的置信区间. T 的分布函数为

$$G(t,p)=P_p(T\leqslant t)=\sum_{i=1}^{[t]}\binom{n}{i}\cdot p^i\cdot(1-p)^{n-i}$$

其中 $[t]$ 表示 $t(0\leqslant t\leqslant n)$ 的整数部分. 不难验证下列等式成立,

$$\sum_{i=0}^{k}\binom{n}{i}\cdot p^i\cdot(1-p)^{n-i}=\frac{\Gamma(n+1)}{\Gamma(k+1)\cdot\Gamma(n-k)}\cdot\int_p^1 u^k\cdot(1-u)^{n-k-1}\mathrm{d}u,\ k=0,1,\cdots,n-1$$

由此及恒等式 $\sum_{i=0}^{n}\binom{n}{i}\cdot p^i\cdot(1-p)^{n-i}=1$ 可见, T 的分布函数 $G(t,p)$ 是 p 的连续、严格减函数.

由样本 $X=(X_1,\cdots,X_n)$ 算得的 $T=\sum_{i=1}^{n}X_i$ 必定是一个正整数. 令其为 k. 在 $k>0$ 时, 利用定理 4.5, 则 p 的置信水平为 $1-\alpha$ 的置信下限是下列方程的解

$$\sum_{i=0}^{k-1}\binom{n}{i}\cdot p^i\cdot(1-p)^{n-i}=1-\alpha \tag{4.14}$$

也就是下列方程的解

$$\frac{\Gamma(n+1)}{\Gamma(k)\cdot\Gamma(n-k+1)}\cdot\int_0^p u^{k-1}\cdot(1-u)^{n-k}\mathrm{d}u=\alpha$$

令 $Be(x|m,n)$ 表示 $Be(m,n)$ 分布的分布函数, 则上式简化为

$$Be(p|k,n-k+1)=\alpha \tag{4.15}$$

不难验证, 若 $B\sim Be(m,n)$, 则 $F=\dfrac{B}{1-B}\cdot\dfrac{n}{m}\sim F(2m,2n)$. 所以方程 (4.15) 可等价变换为

$$F\left(\frac{p}{1-p}\cdot\frac{n-k+1}{k}\,\middle|\,2k,2(n-k+1)\right)=\alpha$$

其中 $F(x|m,n)$ 表示 $F(m,n)$ 分布的分布函数. 因此 p 的置信水平为 $1-\alpha$ 的置信下限 $\hat{p}_L$ 是下列方程的解

$$\frac{p}{1-p}\cdot\frac{n-k+1}{k}=F_{\alpha}(2k,2(n-k+1))$$

由于

$$F_{\alpha}(2k,2(n-k+1))=\frac{1}{F_{1-\alpha}(2(n-k+1),2k)}$$

则有

$$\hat{p}_L=\frac{k}{k+(n-k+1)\cdot F_{1-\alpha}(2(n-k+1),2k)}$$

在 $k=0$ 时,取 $\hat{p}_L=0$.

类似地,由定理 4.5 知,在 $k<n$ 时,p 的置信水平为 $1-\alpha$ 的置信上限 $\hat{p}_U$ 是下列方程的解

$$\sum_{i=0}^{k}\binom{n}{i}\cdot p^i\cdot(1-p)^{n-i}=\alpha$$

把它与(4.14)式进行比较,不难看到 $\hat{p}_U$ 是下列方程的解

$$\frac{p}{1-p}\cdot\frac{n-k}{k+1}=F_{1-\alpha}(2(k+1),2(n-k))$$

所以

$$\hat{p}_U=\frac{(k+1)\cdot F_{1-\alpha}(2(k+1),2(n-k))}{(n-k)\cdot F_{1-\alpha}(2(k+1),2(n-k))+(k+1)}$$

在 $k=n$ 时,取 $\hat{p}_U=1$.

由上述结果可知,p 的置信水平为 $1-\alpha$ 的置信区间为$[\hat{p}_L,\hat{p}_U]$,在 $0<k<n$ 时,

$$\hat{p}_L=\frac{k}{k+(n-k+1)\cdot F_{1-\alpha_1}(2(n-k+1),2k)}$$

$$\hat{p}_U=\frac{(k+1)\cdot F_{1-\alpha_2}(2(k+1),2(n-k))}{(n-k)\cdot F_{1-\alpha_2}(2(k+1),2(n-k))+(k+1)}$$

这里的 $\alpha_1+\alpha_2=\alpha$,且 $0\leqslant\alpha_1,\alpha_2\leqslant1$. 实用时常取 $\alpha_1=\alpha_2=\dfrac{\alpha}{2}$.

在 $k=0$ 时,取

$$\hat{p}_L=0,\quad \hat{p}_U=\frac{F_{1-\alpha}(2,2n)}{n\cdot F_{1-\alpha}(2,2n)+1}$$

在 $k=n$ 时,取

$$\hat{p}_U = 1, \quad \hat{p}_L = \frac{n}{n + F_{1-\alpha}(2, 2n)}$$

§4.2.4　区间估计和假设检验

假设检验和区间估计这两个统计推断问题看似完全不同，而实际上两者之间有着非常密切的联系．正因为如此，Neyman 才可以将 Neyman 和 Pearson 的假设检验理论的基本思想推广到区间估计．

由参数假设检验问题的水平为 α 的检验，可以得到该参数的置信水平为 $1-\alpha$ 的置信区间（置信限）．反之亦然．例如，设样本 $X=(X_1, \cdots, X_n)$ 来自正态总体 $N(\mu, \sigma^2)$，其中：$-\infty<\mu<\infty, \sigma^2>0$．考虑双边假设检验问题，$H_0: \mu=\mu_0$ 对 $H_1: \mu\neq\mu_0$．我们知道，它的水平为 α 的 UMPU 检验的拒绝域为

$$W = \left\{x: |\bar{x} - \mu_0| > \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}(n-1)\right\} \tag{4.16}$$

必须指出的是，由于这里的随机变量是连续型的，所以(4.16)式中的不等号"$>$"可以换为"$\geqslant$"．使用"$\geqslant$"是为了用闭区间给出区间估计．我们在 $|\bar{x}-\mu_0| \leqslant \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}(n-1)$，也就是在

$$\bar{x} - \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}(n-1) \leqslant \mu_0 \leqslant \bar{x} + \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}(n-1)$$

的时候接受原假设．从而有

$$P_{\mu_0}\left\{\bar{X} - \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}(n-1) \leqslant \mu_0 \leqslant \bar{X} + \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}(n-1)\right\}$$

$$= 1 - P_{\mu_0}\left\{|\bar{X} - \mu_0| > \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}(n-1)\right\} = 1 - \alpha$$

考虑到 μ_0 的任意性，则

$$\left[\bar{X} - \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}(n-1), \bar{X} + \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}(n-1)\right]$$

是 μ 的置信水平为 $1-\alpha$ 的置信区间．

如果考虑单边假设检验问题，$H_0: \eta\leqslant\eta_0$ 对 $H_1: \eta>\eta_0$，或 $H_0: \eta\geqslant\eta_0$ 对 $H_1: \eta<\eta_0$，则能得到置信下限，或上限．同上，设样本 $X=(X_1,$

$\cdots,X_n$)来自正态总体 $N(\mu,\sigma^2)$,其中:$-\infty<\mu<\infty,\sigma^2>0$. 考虑单边假设检验问题,$H_0:\mu\leqslant\mu_0$ 对 $H_1:\mu>\mu_0$. 我们知道,水平为 α 的 UMP 检验的拒绝域为

$$W=\left\{x:\bar{x}>\mu_0+\frac{s}{\sqrt{n}}\cdot t_{1-\alpha}(n-1)\right\} \tag{4.17}$$

所以我们在 $\bar{x}\leqslant\mu_0+\frac{s}{\sqrt{n}}\cdot t_{1-\alpha}(n-1)$,也就是在 $\mu_0\geqslant\bar{x}-\frac{s}{\sqrt{n}}\cdot t_{1-\alpha}(n-1)$ 的时候接受原假设. 从而有

$$P_{\mu_0}\left\{\mu_0\geqslant\overline{X}-\frac{s}{\sqrt{n}}\cdot t_{1-\alpha}(n-1)\right\}=1-\alpha$$

考虑到 μ_0 的任意性,则 $\overline{X}-\frac{s}{\sqrt{n}}\cdot t_{1-\alpha}(n-1)$是 μ 的置信水平为 $1-\alpha$ 的置信下限. 同理,如果考虑单边假设检验问题,$H_0:\eta\geqslant\eta_0$ 对 $H_1:\eta<\eta_0$,则能得到置信上限.

由(4.16)和(4.17)两式可以看到,检验的拒绝域 W 不仅与样本观察值 x 有关,还与 μ_0 有关. 故可把 W 记为 $W=W(x,\mu_0)$. 我们在 $x\notin W(x,\mu_0)$的时候接受原假设. μ 的置信区间,或置信下(上)限是在把"$x\notin W(x,\mu_0)$"等价地变换成"$\hat{\mu}_L(x)<\mu_0<\hat{\mu}_U(x)$",或"$\mu_0>\hat{\mu}_L(x)$"("$\mu_0<\hat{\mu}_U(x)$")后得到的. 反之,若能获得某参数的置信限,或置信区间,就可获得该参数的单边,或双边检验问题的拒绝域. 关于区间估计和假设检验的关系的进一步的讨论见下一节.

§4.2.5 似然置信域

设样本 $X=(X_1,\cdots,X_n)$来自密度函数族$\{p(x;\theta):\theta\in\Theta\}$,其中未知参数 $\theta=(\theta_1,\cdots,\theta_k)$,参数空间 Θ 是 k 维欧氏空间 $\mathbf{R}^k$ 中的一个含有内点的集合. 考虑简单原假设 $H_0:\theta=\theta_0$ 对备择假设 $H_1:\theta\neq\theta_0$ 的检验问题. 定理 3.18 告诉我们,在一定的条件下,该检验问题似然比检验的拒绝域为 $W=\{x:2\ln\lambda(x)\dot{>}\chi^2_{1-\alpha}(k)\}$,其中

$$\lambda(X)=\frac{\prod_{i=1}^{n}p(X_i;\hat{\theta})}{\prod_{i=1}^{n}p(X_i;\theta_0)}$$

为似然比统计量，而 $\hat{\theta}$ 是 θ 的 MLE. 显然 $\lambda(X)$ 和 θ_0 有关. 故不妨记 $\lambda(X)$ 为 $\lambda(X;\theta_0)$. 那么集合

$$\{\theta: 2\ln\lambda(x;\theta) \leqslant \chi^2_{1-\alpha}(k)\} = \left\{\theta: 2\ln\left\{\frac{\prod_{i=1}^{n} p(x_i;\hat{\theta})}{\prod_{i=1}^{n} p(x_i;\theta)}\right\} \leqslant \chi^2_{1-\alpha}(k)\right\}$$

就是 $\theta=(\theta_1,\cdots,\theta_k)$ 的置信水平为 $1-\alpha$ 的置信域，通常称为似然置信域.

看习题 3.20 将人按性别与色盲分类，则按遗传学模型，有下列概率模型

	正常	色盲
男	$\frac{p}{2}$	$\frac{1-p}{2}$
女	$\frac{p^2}{2}+p\cdot(1-p)$	$\frac{(1-p)^2}{2}$

其中 $0\leqslant p\leqslant 1$. 今随机调查 1000 人，数据如下

	正常	色盲
男	442	38
女	514	6

试求 p 的置信水平为 95% 的区间估计.

在该概率模型下，似然函数

$$L(p) \propto \left(\frac{p}{2}\right)^{442} \cdot \left(\frac{p(2-p)}{2}\right)^{514} \cdot \left(\frac{1-p}{2}\right)^{38} \cdot \left(\frac{(1-p)^2}{2}\right)^{6}$$

$$= p^{956} \cdot (2-p)^{514} \cdot (1-p)^{50}$$

且 p 的 MLE 为 $\hat{p}=0.913$. 故下列集合为 p 的置信水平为 95% 的似然置信区间

$$\left\{p : 2\ln\frac{L(\hat{p})}{L(p)} \leqslant \chi^2_{0.95}(1) = 3.841\right\}$$

从而有不等式

$$956\ln p + 514\ln(2-p) + 50\ln(1-p) \geqslant -168.1487$$

令

$$L(p) = 956\ln p + 514\ln(2-p) + 50\ln(1-p)$$

由于

$$L'(p) = \frac{956}{p} - \frac{514}{2-p} - \frac{50}{1-p} = \frac{1520 \cdot p^2 - 3482 \cdot p + 1912}{p \cdot (2-p) \cdot (1-p)}$$

所以在 $p \in [0,1]$时,$L(p)$是下凸函数.因为

$$L(0.8899) = -168.1447 \qquad L(0.9331) = -168.1388$$

$$L(0.8898) = -168.1605 \qquad L(0.9332) = -168.1593$$

所以不等式的解为 $0.890 \leqslant p \leqslant 0.933$. 故 p 的置信水平为 95%的似然置信区间为[0.890,0.933].

似然置信区间方法的应用范围很广,它可用于很一般的场合.

§ 4.3 一致最精确的置信区间(置信限)

§ 4.3.1 一致最精确的置信限

在§4.1我们知道,若参数真值为 θ,则对置信下限 $\hat{\theta}_L(X)$来说,在 $\theta' < \theta$ 时,θ' 大于或等于 $\hat{\theta}_L(X)$的概率 $P_\theta\{\theta' \geqslant \hat{\theta}_L(X)\}$愈小,置信下限的精度愈高.同样地,对置信上限 $\hat{\theta}_U(X)$来说,在 $\theta' > \theta$ 时,θ' 小于或等于 $\hat{\theta}_U(X)$的概率 $P_\theta\{\theta' \leqslant \hat{\theta}_U(X)\}$愈小,置信上限的精度愈高.据此有如下的定义.

定义 4.6 设 $\hat{\theta}_L(X)$是 θ 的置信水平为 $1-\alpha$ 的置信下限.如果对任意一个 θ 的置信水平为 $1-\alpha$ 的置信下限 $\hat{\theta}_L^*(X)$,在 $\theta' < \theta$ 时,都有

$$P_\theta\{\theta' \geqslant \hat{\theta}_L(X)\} \leqslant P_\theta\{\theta' \geqslant \hat{\theta}_L^*(X)\}$$

则称 $\hat{\theta}_L(X)$是 θ 的置信水平为 $1-\alpha$ 的一致最精确(Uniformly Most Accurate,简称 UMA)置信下限.

定义 4.7 设 $\hat{\theta}_U(X)$是 θ 的置信水平为 $1-\alpha$ 的置信上限.如果对任意一个 θ 的置信水平为 $1-\alpha$ 的置信上限 $\hat{\theta}_U^*(X)$,在 $\theta'>\theta$ 时,都有

$$P_\theta\{\theta' \leqslant \hat{\theta}_U(X)\} \leqslant P_\theta\{\theta' \leqslant \hat{\theta}_U^*(X)\}$$

则称 $\hat{\theta}_U(X)$是 θ 的置信水平为 $1-\alpha$ 的 UMA 置信上限.

在 §4.2.4 我们知道,由单边假设检验问题可以得到置信限.不仅如此,由单边假设检验问题的 UMP 检验可以得到 UMA 置信限.

定理 4.7 设单边假设检验问题,$H_0:\theta\leqslant\theta_0$ 对 $H_1:\theta>\theta_0$ 的水平为 α 的 UMP 检验的拒绝域为 $W=W(x;\theta_0)$.则检验的接受域为 $\overline{W}=\overline{W}(x;\theta_0)$.如果"$x\in\overline{W}(x;\theta_0)$"可以等价地变换为"$\theta_0\geqslant\hat{\theta}_L(X)$",那么 $\hat{\theta}_L(X)$是 θ 的置信水平为 $1-\alpha$ 的 UMA 置信下限.

证明:由于 $W=W(x;\theta_0)$是所考虑的单边假设检验问题的水平为 α 的检验的拒绝域,所以

$$\begin{aligned}P_{\theta_0}\{\theta_0 \geqslant \hat{\theta}_L(X)\} &= P_{\theta_0}\{X \in \overline{W}(X;\theta_0)\}\\ &= 1 - P_{\theta_0}\{X \in W(X;\theta_0)\} \geqslant 1-\alpha\end{aligned}$$

考虑到 θ_0 的任意性,故 $\hat{\theta}_L(X)$是 θ 的置信水平为 $1-\alpha$ 的置信下限.

对任意一个 θ 的置信水平为 $1-\alpha$ 的置信下限 $\hat{\theta}_L^*(X)$,构造检验,令其拒绝域为

$$W^* = W^*(x;\theta_0) = \{x:\theta_0 < \hat{\theta}_L^*(X)\}$$

则在 $\theta\leqslant\theta_0$ 时,

$$\begin{aligned}P_\theta\{X \in W^*\} &= P_\theta\{\theta_0 < \hat{\theta}_L^*(X)\} \leqslant P_\theta\{\theta < \hat{\theta}_L^*(X)\}\\ &= 1 - P_\theta\{\theta \geqslant \hat{\theta}_L^*(X)\} \leqslant 1-(1-\alpha) = \alpha\end{aligned}$$

所以拒绝域为 $W^*=W^*(x;\theta_0)$的检验是所考虑的单边假设检验问题的水平为 α 的一个检验.由于拒绝域为 $W=W(x;\theta_0)$的检验是所考虑的单边假设检验问题的水平为 α 的 UMP 检验.所以在 $\theta>\theta_0$ 时,

$$\begin{aligned}&P_\theta\{X \in W(X;\theta_0)\} \geqslant P_\theta\{X \in W^*(X;\theta_0)\}\\ &\Rightarrow P_\theta\{\theta_0 < \hat{\theta}_L(X)\} \geqslant P_\theta\{\theta_0 < \hat{\theta}_L^*(X)\}\\ &\Rightarrow P_\theta\{\theta_0 \geqslant \hat{\theta}_L(X)\} \leqslant P_\theta\{\theta_0 \geqslant \hat{\theta}_L^*(X)\}\end{aligned}$$

这说明 $\hat{\theta}_L(X)$是 θ 的置信水平为 $1-\alpha$ 的 UMA 置信下限.证毕.

关于 UMA 置信上限有类似的结果.

定理 4.8 设单边假设检验问题,$H_0: \theta \geqslant \theta_0$ 对 $H_1: \theta < \theta_0$ 的水平为 α 的 UMP 检验的拒绝域为 $W = W(x;\theta_0)$,接受域为 $\overline{W} = \overline{W}(x;\theta_0)$.如果"$x \in \overline{W}(x;\theta_0)$"可以等价地变换为"$\theta_0 \leqslant \hat{\theta}_U(X)$",那么 $\hat{\theta}_U(X)$是 θ 的置信水平为 $1-\alpha$ 的 UMA 置信上限.

§ 4.3.2 一致最精确的无偏置信限和无偏置信区间

定理 4.7 和 4.8 告诉我们,UMA 置信限可由单参数假设检验问题的 UMP 检验导出.一般来说,在多参数分布族的场合下,单参数假设检验问题的 UMP 检验不存在,因而 UMA 置信限也不存在.多参数分布族场合,单参数假设检验问题的 UMPU 检验常是存在的,由它可导出一致最精确无偏(Uniformly Most Accurate Unbiased,简称 UMAU)置信限.无偏置信限和 UMAU 置信限的定义如下.

定义 4.8 设 $\hat{\theta}_L(X)$是 θ 的置信水平为 $1-\alpha$ 的置信下限.若在 $\theta' < \theta$ 时,都有

$$P_\theta\{\theta' \geqslant \hat{\theta}_L(X)\} \leqslant 1-\alpha$$

则称 $\hat{\theta}_L(X)$是 θ 的置信水平为 $1-\alpha$ 的无偏置信下限.若对任意一个 θ 的置信水平为 $1-\alpha$ 的无偏置信下限 $\hat{\theta}_L^*(X)$,在 $\theta' < \theta$ 时,都有

$$P_\theta\{\theta' \geqslant \hat{\theta}_L(X)\} \leqslant P_\theta\{\theta' \geqslant \hat{\theta}_L^*(X)\}$$

则称 $\hat{\theta}_L(X)$是 θ 的置信水平为 $1-\alpha$ 的 UMAU 置信下限.

定义 4.9 设 $\hat{\theta}_U(X)$是 θ 的置信水平为 $1-\alpha$ 的置信上限.如果在 $\theta' > \theta$ 时,都有

$$P_\theta\{\theta' \leqslant \hat{\theta}_U(X)\} \leqslant 1-\alpha$$

则称 $\hat{\theta}_U(X)$是 θ 的置信水平为 $1-\alpha$ 的无偏置信上限.如果对任意一个 θ 的置信水平为 $1-\alpha$ 的无偏置信上限 $\hat{\theta}_U^*(X)$,在 $\theta' > \theta$ 时,都有

$$P_\theta\{\theta' \leqslant \hat{\theta}_U(X)\} \leqslant P_\theta\{\theta' \leqslant \hat{\theta}_U^*(X)\}$$

则称 $\hat{\theta}_U(X)$是 θ 的置信水平为 $1-\alpha$ 的 UMAU 置信上限.

在 § 4.2 我们知道,由双边假设检验问题 $H_0: \eta = \eta_0$ 对 $H_1: \eta \neq \eta_0$ 的水平为 α 的检验可以得到 η 的置信水平为 $1-\alpha$ 的置信区间.一般来说,在单参数分布族或多参数分布族的情况,这类双边假设检验问题不存在 UMP 检验,而仅存在 UMPU 检验.由 UMPU 检验可导出

UMAU置信区间. 无偏置信区间和 UMAU 置信区间的定义如下.

定义 4.10　设$[\hat{\theta}_L(X),\hat{\theta}_U(X)]$是 θ 的置信水平为 $1-\alpha$ 的置信区间. 如果在 $\theta'\neq\theta$ 时，都有

$$P_\theta\{\hat{\theta}_L(X)\leqslant\theta'\leqslant\hat{\theta}_U(X)\}\leqslant 1-\alpha$$

则称$[\hat{\theta}_L(X),\hat{\theta}_U(X)]$是 θ 的置信水平为 $1-\alpha$ 的无偏置信区间. 如果对任意一个 θ 的置信水平为 $1-\alpha$ 的无偏置信区间$[\hat{\theta}_L^*(X),\hat{\theta}_U^*(X)]$，在 $\theta'\neq\theta$ 时，都有

$$P_\theta\{\hat{\theta}_L(X)\leqslant\theta'\leqslant\hat{\theta}_U(X)\}\leqslant P_\theta\{\hat{\theta}_L^*(X)\leqslant\theta'\leqslant\hat{\theta}_U^*(X)\}$$

则称$[\hat{\theta}_L(X),\hat{\theta}_U(X)]$是 θ 的置信水平为 $1-\alpha$ 的 UMAU 置信区间.

定理 4.9　设双边假设检验问题 $H_0:\theta=\theta_0$ 对 $H_1:\theta\neq\theta_0$ 的水平为 α 的 UMPU 检验的拒绝域为 $W=W(x;\theta_0)$，接受域为 $\overline{W}=\overline{W}(x;\theta_0)$. 如果"$x\in\overline{W}(x;\theta_0)$"可以等价地变换为"$\hat{\theta}_L(X)\leqslant\theta_0\leqslant\hat{\theta}_U(X)$"，那么$[\hat{\theta}_L(X),\hat{\theta}_U(X)]$是 θ 的置信水平为 $1-\alpha$ 的 UMAU 置信区间.

证明：由于 $W=W(x;\theta_0)$是所考虑的双边假设检验问题的水平为 α 的无偏检验的拒绝域，所以

$$\begin{aligned}P_{\theta_0}\{\hat{\theta}_L(X)\leqslant\theta_0\leqslant\hat{\theta}_L(X)\}&=P_{\theta_0}\{X\in\overline{W}(X;\theta_0)\}\\&=1-P_{\theta_0}\{X\in W(X;\theta_0)\}\geqslant 1-\alpha\end{aligned}$$

并且在 $\theta\neq\theta_0$ 时，

$$\begin{aligned}P_{\theta_0}\{\hat{\theta}_L(X)\leqslant\theta\leqslant\hat{\theta}_L(X)\}&=P_{\theta_0}\{X\in\overline{W}(X;\theta)\}\\&=1-P_{\theta_0}\{X\in W(X;\theta)\}\leqslant 1-\alpha\end{aligned}$$

故$[\hat{\theta}_L(X),\hat{\theta}_U(X)]$是 θ 的置信水平为 $1-\alpha$ 的无偏置信区间.

对任意一个 θ 的置信水平为 $1-\alpha$ 的无偏置信区间$[\hat{\theta}_L^*(X),\hat{\theta}_U^*(X)]$，构造检验，使其拒绝域为

$$W^*=W^*(X;\theta_0)=\{x;\theta_0<\hat{\theta}_L^*(x)\text{ 或 }\theta_0>\hat{\theta}_U^*(x)\}$$

则

$$\begin{aligned}P_{\theta_0}\{X\in W^*\}=1-P_{\theta_0}\{X\in\overline{W}^*\}&=1-P_{\theta_0}\{\hat{\theta}_L^*(X)\\&\leqslant\theta_0\leqslant\hat{\theta}_U^*(X)\}\leqslant 1-(1-\alpha)=\alpha\end{aligned}$$

并且在 $\theta\neq\theta_0$ 时，

$$P_\theta\{X\in W^*\}=1-P_\theta\{X\in\overline{W}^*\}=1-P_\theta\{\hat{\theta}_L^*(X)$$

$$\leqslant \theta_0 \leqslant \hat{\theta}_U^*(X)\} \geqslant 1-(1-\alpha)=\alpha$$

所以拒绝域为 $W^*=W^*(x;\theta_0)$ 的检验是所考虑的双边假设检验问题的水平为 α 的一个无偏检验. 由于拒绝域为 $W=W(x;\theta_0)$ 的检验是所考虑的双边假设检验问题的水平为 α 的 UMPU 检验. 所以在 $\theta\neq\theta_0$ 时,

$$P_\theta\{X\in W(X;\theta_0)\}\geqslant P_\theta\{X\in W^*(X;\theta_0)\}$$

$$\Rightarrow P_\theta\{X\in \overline{W}(X;\theta_0)\}\leqslant P_\theta\{X\in \overline{W}^*(X;\theta_0)\}$$

$$\Rightarrow P_\theta\{\hat{\theta}_L(X)\leqslant\theta_0\leqslant\hat{\theta}_U(X)\}\leqslant P_\theta\{\hat{\theta}_L^*(X)\leqslant\theta_0\leqslant\hat{\theta}_U^*(X)\}$$

这说明$[\hat{\theta}_L(X),\hat{\theta}_U(X)]$是 θ 的置信水平为$1-\alpha$的 UMAU 置信区间. 证毕.

关于 UMAU 置信限有类似的结果.

定理 4.10 设单边假设检验问题,$H_0:\theta\leqslant\theta_0$ 对 $H_1:\theta>\theta_0$ 的水平为 α 的 UMPU 检验的拒绝域为 $W=W(x;\theta_0)$,接受域为 $\overline{W}=\overline{W}(x;\theta_0)$. 如果"$x\in\overline{W}(x;\theta_0)$"可以等价地变换为"$\theta_0\geqslant\hat{\theta}_L(X)$",那么 $\hat{\theta}_L(X)$是 θ 的置信水平为 $1-\alpha$ 的 UMAU 置信下限.

定理 4.11 设单边假设检验问题,$H_0:\theta\geqslant\theta_0$ 对 $H_1:\theta<\theta_0$ 的水平为 α 的 UMPU 检验的拒绝域为 $W=W(x;\theta_0)$,接受域为 $\overline{W}=\overline{W}(x;\theta_0)$. 如果"$x\in\overline{W}(x;\theta_0)$"可以等价地变换为"$\theta_0\leqslant\hat{\theta}_U(x)$",那么 $\hat{\theta}_U(X)$是 θ 的置信水平为 $1-\alpha$ 的 UMAU 置信上限.

例 4.9 设总体为$\{N(\mu,\sigma^2):-\infty<\mu<\infty,\sigma^2>0\}$. 又设 $X=(X_1,\cdots,X_n)$是来自该总体的样本. 由例 3.11 知,检验问题 $H_0:\sigma^2=\sigma_0^2$ 对 $H_1:\sigma^2\neq\sigma_0^2$ 的水平为 $\alpha(0<\alpha<1)$的 UMPU 检验 $\phi(t)$的拒绝域为 $W=\{x:T(x)\leqslant c_1$ 或 $T(x)\geqslant c_2\}$,其中 c_1 和 c_2 由下式确定,

$$c_2\cdot\chi^2\left(\frac{c_2}{\sigma_2^2}\,\middle|\,n-1\right)-c_1\cdot\chi^2\left(\frac{c_1}{\sigma_0^2}\,\middle|\,n-1\right)=0$$

其中 $T(x)=\sum_{i=1}^n(x_i-\overline{x})^2$,$\chi^2(x\mid n-1)$表示自由度为 $n-1$ 的 χ^2 分布的密度函数. 令 $c_1=\frac{(n-1)\cdot\sigma_0^2}{b}$,$c_2=\frac{(n-1)\cdot\sigma_0^2}{a}$. 则拒绝域可等价地改写为 $W=\{x:b\cdot S^2\leqslant\sigma_0^2$ 或 $a\cdot S^2\geqslant\sigma_0^2\}$. 其中 $S^2=\frac{\sum_{i=1}^n(x_i-\overline{x})^2}{n-1}=$

$\frac{T(x)}{n-1}$，a 和 b 满足下列条件

$$\frac{1}{a}\cdot\chi^2\left(\frac{n-1}{a}\middle|n-1\right)=\frac{1}{b}\cdot\chi^2\left(\frac{n-1}{b}\middle|n-1\right) \tag{4.18}$$

我们不妨把拒绝域写为 $W=\{x: b\cdot S^2<\sigma_0^2$ 或 $a\cdot S^2>\sigma_0^2\}$，则由定理 4.9 知，$[a\cdot S^2, b\cdot S^2]$ 为 σ^2 的置信水平为 $1-\alpha$ 的 UMAU 置信区间.

由例 4.2 知，在形如 $[a\cdot S^2, b\cdot S^2]$ 的区间估计中，当 a 和 b 满足 (4.4) 式，即

$$\frac{1}{a^2}\cdot\chi^2\left(\frac{n-1}{a}\middle|n-1\right)=\frac{1}{b^2}\cdot\chi^2\left(\frac{n-1}{b}\middle|n-1\right)$$

的时候，$[a\cdot S^2, b\cdot S^2]$ 是 σ^2 的置信水平为 $1-\alpha$ 的区间平均长度最短的置信区间. 显然，(4.18) 式和 (4.4) 式不等价. 所以 σ^2 的置信水平为 $1-\alpha$ 的 UMAU 置信区间的平均长度并不是最短的.

§ 4.3.3　置信区间的平均长度

在 § 4.1 我们知道，区间估计的精确度的标准有两个. 一个标准是，区间包含非真值的概率愈小愈好. UMA 和 UMAU 就是根据这个标准而提出的. 另一个标准是，区间的平均长度愈短愈好. 下面的定理告诉我们，这两个标准之间有着某种联系.

定理 4.12　设参数空间 $\Theta=\{\theta\}$，Θ 是直线上的含有内点的一个区间. 如果 $[\hat\theta_L(X), \hat\theta_U(X)]$ 是参数 θ 的置信水平为 $1-\alpha$ 的 UMAU 置信区间. 则对 θ 的任意一个置信水平为 $1-\alpha$ 的无偏置信区间 $[\hat\theta_L^*(x), \hat\theta_U^*(x)]$，都有

$$\mathrm{E}_\theta[\hat\theta_U(X)-\hat\theta_L(X)]\leqslant \mathrm{E}_\theta[\hat\theta_U^*(X)-\hat\theta_L^*(X)]$$

对一切的 θ 都成立.

证明：由定理的条件知

$$\begin{aligned}\mathrm{E}_\theta[\hat\theta_U(X)-\hat\theta_L(X)] &= \iint_{\hat\theta_L(x)}^{\hat\theta_U(x)}\mathrm{d}\theta'\mathrm{d}P_\theta(x)\\ &= \iint\limits_{\{x:\ \hat\theta_L(x)\leqslant\theta'\leqslant\hat\theta_U(x)\}}\mathrm{d}P_\theta(x)\mathrm{d}\theta'\end{aligned}$$

$$= \int P_{\theta}\{\hat{\theta}_L(X) \leqslant \theta' \leqslant \hat{\theta}_U(X)\}\mathrm{d}\theta'$$

$$= \int_{\theta' \neq \theta} P_{\theta}\{\hat{\theta}_L(X) \leqslant \theta' \leqslant \hat{\theta}_U(X)\}\mathrm{d}\theta'$$

$$\leqslant \int_{\theta' \neq \theta} P_{\theta}\{\hat{\theta}_L^*(X) \leqslant \theta' \leqslant \hat{\theta}_U^*(X)\}\mathrm{d}\theta'$$

$$= \mathrm{E}_{\theta}[\hat{\theta}_U^*(X) - \hat{\theta}_L^*(X)]$$

定理证毕.

在例 4.10 中,若仅考虑 σ^2 的置信水平为 $1-\alpha$ 的无偏置信区间,则由定理 4.12 知,σ^2 的置信水平为 $1-\alpha$ 的 UMAU 置信区间的平均长度最短.但正如前面所说的,在置信水平为 $1-\alpha$ 的置信区间中,它并不是区间平均长度最短的.

关于置信限,类似于定理 4.12 的结论不一定成立.看下面的例子.

例 4.10 设总体服从指数分布,其密度函数为

$$p(x;\theta) = \begin{cases} \dfrac{1}{\theta} \cdot \exp\left\{-\dfrac{x}{\theta}\right\}, & x \geqslant 0 \\ 0, & x < 0 \end{cases}$$

$X_1,\cdots,X_n$ 为来自该总体的样本.由定理 4.8 知,θ 的置信水平为 $1-\alpha$ 的 UMA 置信上限为 $\hat{\theta}_U = 2\sum_{i=1}^n X_i/\chi_\alpha^2(2n)$.显然,$\hat{\theta}_U$ 是 θ 的置信水平为 $1-\alpha$ 的 UMAU 置信上限.不难算得,$\mathrm{E}_\theta(\hat{\theta}_U) = 2n\theta/\chi_\alpha^2(2n)$.

令 $X_{(1)}$ 为最小次序统计量.由于 $X_{(1)}$ 的密度函数为

$$p(y;\theta) = \begin{cases} \dfrac{n}{\theta} \cdot \exp\left\{-\dfrac{ny}{\theta}\right\}, & y \geqslant 0 \\ 0, & y < 0 \end{cases}$$

所以 $\hat{\theta}_U^* = 2nX_{(1)}/\chi_\alpha^2(2)$ 为 θ 的置信水平为 $1-\alpha$ 的置信上限.可以证明,$\hat{\theta}_U^*$ 为 θ 的置信水平为 $1-\alpha$ 的无偏置信上限(习题 4.11).不难算得,$\mathrm{E}_\theta(\hat{\theta}_U^*) = \dfrac{2\theta}{\chi_\alpha^2(2)}$.

取 $\alpha = 0.7$,$n=6$.由于

$$\mathrm{E}_\theta(\hat{\theta}_U^*) = \frac{2\theta}{\chi_{0.7}^2(2)} = \frac{2\theta}{2.408} = 0.831\theta$$

$$E_\theta(\hat{\theta}_U) = \frac{12\theta}{\chi^2_{0.7}(12)} = \frac{12\theta}{14.011} = 0.857\theta$$

所以 $E_\theta(\hat{\theta}_U^*) < E_\theta(\hat{\theta}_U)$. 这样,在平均的意义下一个不是 UMAU 的无偏置信上限小于具有相同置信系数的 UMAU 置信上限.

§4.4　信仰推断方法

§4.4.1　信仰分布

参数估计中的极大似然方法最早是由 C. F. Gauss 在 1821 年提出的. 然而这个方法应主要归功于 R. A. Fisher. 他在 1922 年重新提出了这个方法,并首先研究了这个方法的统计性质. 设样本 $X \sim p(x;\theta)$, $\theta \in \Theta$. 在有了样本观察值 x 后,如果对不同的 $\theta_1, \theta_2 \in \Theta$,有 $p(x;\theta_1) > p(x;\theta_2)$,那么观察值 x 来自总体 $p(x;\theta_1)$ 比来自总体 $p(x;\theta_2)$ 的可能性大,也就是说 $\theta=\theta_1$ 比 $\theta=\theta_2$ 的可能性大. 据此有 θ 的 MLE $\hat{\theta}=\hat{\theta}(x)$,它满足条件 $p(x;\hat{\theta}) = \sup_{\theta\in\Theta} p(x;\theta)$. 所以 Fisher 的极大似然估计的基本想法是,在有了样本观察值 x 后,用"最有可能出现"的 θ 去估计 θ. Fisher 在本世纪三十年代进一步推广了极大似然的想法,提出了信仰推断方法. 他认为在有了样本观察值 x 后,就有了参数 θ 的一个分布. 这个分布表示了,由于所得样本观察值的信息,θ 落在各个范围内的"可信程度". Fisher 称这个分布为信仰分布(Fiducial Distribution). 看下面的例子.

例 4.11　设样本 $X=(X_1,\cdots,X_n)$ 来自正态总体 $N(\mu,1)$,其中 $-\infty<\mu<\infty$. 由于 $\overline{X}$ 是 μ 的充分统计量,所以我们基于 $\overline{X}$ 来考虑该问题. 因为 $\overline{X} \sim N(\mu, \frac{1}{n})$,故若记 $e \sim N(0,1)$,则有

$$\overline{X} = \mu + \frac{1}{\sqrt{n}} \cdot e \tag{4.19}$$

移项可得

$$\mu = \overline{X} - \frac{1}{\sqrt{n}} \cdot e \tag{4.20}$$

(4.20)式由(4.19)式等价变换而来，但 Fisher 给予(4.20)式以完全新的解释. 他认为在有了样本观察值 X，并继而得到 $\overline{X}$ 后，(4.20)式给出了 μ 的分布，$\mu\sim N(\overline{X},\frac{1}{n})$. Fisher 称这个分布为 μ 的信仰分布. 由 μ 的这个信仰分布可以得到

$$P\left\{\overline{X}-\frac{1}{\sqrt{n}}\cdot U_{1-\frac{\alpha}{2}}\leqslant\mu\leqslant\overline{X}+\frac{1}{\sqrt{n}}\cdot U_{1-\frac{\alpha}{2}}\right\}=1-\alpha$$

这里的 $1-\alpha$ 称为信仰系数，$\left[\overline{X}-\frac{1}{\sqrt{n}}\cdot U_{1-\frac{\alpha}{2}},X+\frac{1}{\sqrt{n}}\cdot U_{1-\frac{\alpha}{2}}\right]$称为 μ 的信仰水平为 $1-\alpha$ 的区间估计. 这个区间也是 μ 的置信水平为 $1-\alpha$ 的置信区间. 有的时候信仰水平为 $1-\alpha$ 的区间估计是置信水平为 $1-\alpha$ 的区间估计，如本例所示. 但有的时候信仰水平为 $1-\alpha$ 的区间估计并不是置信水平为 $1-\alpha$ 的区间估计，如§4.4.3 的 Behrens-Fisher 问题所示.

§4.4.2 函数模型

关于信仰分布的诱导，有好几种方法. 建立函数模型诱导信仰分布是较为简单的一种方法.

定义 4.11 函数模型有三个要素，观察值 X，取值于$(\mathscr{X},\mathscr{B}_{\mathscr{X}})$；参数 θ，取值于$(\Theta,\mathscr{B}_{\Theta})$；误差变量 e，取值于$(\mathscr{E},\mathscr{B}_{e})$，$e$ 的分布 P 与 θ 无关. 在这三个量之间存在一个函数关系：$(\theta,e)\to X$，记为 $X=\theta\circ e$，其中"$\circ$"表示某种运算. 要求对每一个 $X\in\mathscr{X}$，存在 θ 和 e，使得 $X=\theta\circ e$. 称"$X=\theta\circ e,e\sim P$"为函数模型. 函数模型常简记为"$X=\theta\circ e$".

例 4.11 中的(4.19)式就是一个函数模型.

在函数模型中，当 θ 给定时，X 是 e 的函数. 则由"$X=\theta\circ e,e\sim P$"可以推得 X 的分布：$X\sim P_\theta(x)$. 所以在 θ 给定时由函数模型"$X=\theta\circ e,e\sim P$"诱导出一个统计结构：$X\sim P_\theta(x),\theta\in\Theta$.

此外，如果对所有的 X 和 e，存在唯一的 θ 使得 $X=\theta\circ e$，则在有了样本观察值 X 后，θ 是 e 的函数. 从而由函数模型"$X=\theta\circ e,e\sim P$"诱导出 θ 的分布. 这就是 θ 的信仰分布.

例 4.12 设样本 $X=(X_1,\cdots,X_n)$来自均匀分布总体 $R(0,\theta)$，其

中 $\theta>0$. 由于最大次序统计量 $X_{(n)}$ 是 θ 的充分统计量. 所以我们基于 $X_{(n)}$ 来考虑该问题. 因为 $X_{(n)}$ 的密度函数为 $p(y;\theta)=\frac{ny^{n-1}}{\theta^n}, 0\leqslant y\leqslant\theta$, 令 $e=\frac{X_{(n)}}{\theta}$. 则 e 的密度函数为 $p(t)=nt^{n-1}, 0\leqslant t\leqslant 1$. 故 e 分布与 θ 无关. 从而有函数模型

$$X_{(n)}=\theta\cdot e \tag{4.21}$$

移项可得 $\theta=\frac{X_{(n)}}{e}$. 则在有了样本观察值 X, 从而有了 $X_{(n)}$ 后, 由此函数模型诱导出 θ 的信仰分布. θ 的信仰分布的密度函数为 $\frac{n\cdot X_{(n)}^n}{\theta^{n+1}}, \theta\geqslant X_{(n)}$. 下面我们根据 θ 的这个信仰分布构造 θ 的区间估计. 由于该信仰分布的密度函数关于 θ 严格单调下降, 所以从区间长度尽可能短的精度要求出发, 我们取形如 $[X_{(n)}, \hat{\theta}_U]$ 的区间为 θ 的区间估计. 从而由

$$P\{X_{(n)}\leqslant\theta\leqslant\hat{\theta}_U\}=\int_{X_{(n)}}^{\hat{\theta}_U}\frac{n\cdot X_{(n)}^n}{\theta^{n+1}}\mathrm{d}\theta=1-\alpha$$

推得 $\hat{\theta}_U=\frac{X_{(n)}}{\sqrt[n]{\alpha}}$. 所以 θ 的信仰水平为 $1-\alpha$ 的区间估计为 $\left[X_{(n)}, \frac{X_{(n)}}{\sqrt[n]{\alpha}}\right]$. 它与例 4.3 给出的 θ 的置信水平为 $1-\alpha$ 的置信区间是一致的.

例 4.13 设样本 $X=(X_1,\cdots,X_n)$ 来自正态分布总体 $N(\mu,\sigma^2)$, 其中 $-\infty<\mu<\infty, \sigma^2>0$. 由于 $(\overline{X}, Q^2=\sum_{i=1}^n(X_i-\overline{X})^2)$ 是 (μ,σ^2) 的充分统计量, 所以我们基于 $(\overline{X},Q^2)$ 来考虑该问题. 因为 $\overline{X}\sim N\left(\mu,\frac{\sigma^2}{\sqrt{n}}\right)$, $\frac{Q^2}{\sigma^2}\sim\chi^2(n-1)$, 并且 $\overline{X}$ 和 Q^2 相互独立, 故若记 $e_1\sim N(0,1)$, $e_2\sim\chi^2(n-1)$, 且 e_1 和 e_2 相互独立, 则有函数模型

$$\begin{cases}\overline{X}=\mu+\frac{\sigma}{\sqrt{n}}\cdot e_1\\ Q^2=\sigma^2\cdot e_2\end{cases} \tag{4.22}$$

此函数模型中的观察值, 参数和误差变量分别为 $(\overline{X},Q^2)$, (μ,σ^2) 和 (e_1,e_2).

由此函数模型可得

$$
\begin{cases}
\mu = \bar{X} - \dfrac{1}{\sqrt{n}} \cdot \dfrac{Q}{\sqrt{e_2}} \cdot e_1 \\[2ex]
\sigma^2 = \dfrac{Q^2}{e_2}
\end{cases}
\tag{4.23}
$$

则在有了样本观察值 X,从而有了 $\overline{X}$ 和 Q^2 后,可诱导出(μ,σ^2)的联合信仰分布.由于 $e_1 \sim N(0,1)$,$e_2 \sim \chi^2(n-1)$,且 e_1 和 e_2 相互独立,则(μ,σ^2)的联合信仰密度函数为

$$
p(\mu,\tau) \propto \exp\left\{-\frac{n \cdot (\mu - \overline{X})^2 + Q^2}{2 \cdot \tau}\right\} \cdot \tau^{-\frac{n+2}{2}},
$$

$$
-\infty < \mu < \infty, \tau > 0 \tag{4.24}
$$

其中 $\tau = \sigma^2$. 由此联合信仰分布可分别推导出 μ 和 σ^2 的边际信仰分布.此外,我们也可直接由(4.23)式分别推导出 μ 和 σ^2 的边际信仰分布.由(4.23)式可得

$$
\sqrt{n-1} \cdot \frac{\sqrt{n}\,(\mu - \overline{X})}{Q} = -\sqrt{n-1} \cdot \frac{e_1}{\sqrt{e_2}} \sim t(n-1)
$$

所以 μ 的边际信仰分布为 t 分布.由此信仰分布所构造的 μ 的信仰水平为 $1-\alpha$ 的区间估计与例 4.1 给出的 μ 的置信水平为 $1-\alpha$ 的区间估计相一致(习题 4.15).由(4.23)式可得

$$
\frac{\sigma^2}{Q^2} = \frac{1}{e_2}, e_2 \sim \chi^2(n-1)
$$

所以 σ^2 的边际信仰分布为倒 χ^2 分布,其密度函数为

$$
\propto \exp\left\{-\frac{Q^2}{2 \cdot \tau}\right\} \cdot \tau^{-\frac{n+1}{2}}, \tau > 0
$$

设由此信仰分布所构造的 σ^2 的信仰水平为 $1-\alpha$ 的区间估计为$[\hat{\sigma}_L^2, \hat{\sigma}_U^2]$.根据区间长度尽可能短的精度要求,区间的左、右端点 $\hat{\sigma}_L^2$ 和 $\hat{\sigma}_U^2$ 须满足条件(习题 4.15)

$$
(\hat{\sigma}_L^2)^{-\frac{n+1}{2}} \cdot \exp\left\{-\frac{Q^2}{2 \cdot \hat{\sigma}_L^2}\right\} = (\hat{\sigma}_U^2)^{-\frac{n+1}{2}} \cdot \exp\left\{-\frac{Q^2}{2 \cdot \hat{\sigma}_U^2}\right\}
$$

如果把 $\hat{\sigma}_L^2$ 和 $\hat{\sigma}_U^2$ 分别记为 $a \cdot S^2$ 和 $b \cdot S^2$,其中 $S^2 = \dfrac{\sum\limits_{i=1}^{n}(X_i - \overline{X})^2}{n-1} =$

$\frac{Q^2}{n-1}$. 那么 a 和 b 就满足条件

$$(a)^{-\frac{n+1}{2}} \cdot \exp\left\{-\frac{n-1}{2\cdot a}\right\} = (b)^{-\frac{n+1}{2}} \cdot \exp\left\{-\frac{n-1}{2\cdot b}\right\}$$

所以 σ^2 的信仰水平为 $1-\alpha$ 的区间估计与例 4.2 给出的 σ^2 的置信水平为 $1-\alpha$ 的区间估计相一致.

§4.4.3 Behrens-Fisher 问题

1929 年 Behrens 提出了一个非常实用的问题. 设样本 $X=(X_1,\cdots,X_m)$ 和样本 $Y=(Y_1,\cdots,Y_n)$ 互相独立. 它们分别来自正态总体 $N(\mu_1,\sigma_1^2)$ 和 $N(\mu_2,\sigma_2^2)$, 其中 $-\infty<\mu_1<\infty$, $-\infty<\mu_2<\infty$, $\sigma_1^2>0$, $\sigma_2^2>0$. 要构造 $\mu_1-\mu_2$ 的区间估计. Fisher 曾研究过这个问题, 并用他提出的信仰推断方法给出了一个解法. 所以人们习惯称这个问题为 Behrens-Fisher 问题.

在 $\sigma_1^2=\sigma_2^2=\sigma^2$, 但 σ^2 未知时, 问题已得到解决(见例 4.4). 在 σ_1^2 和 σ_2^2 完全未知的时候, 要想得到一个满意的解答是很困难的. 下面介绍一些较为重要的解法. 1944 年 Scheffe 给出了服从 t 分布且含有参数 $\mu_1-\mu_2$ 的一个枢轴量, 从而得到了 $\mu_1-\mu_2$ 的置信水平为 $1-\alpha$ 的区间估计, 给出了 Behrens-Fisher 问题的一个解法[1]. 若考虑以区间的平均长度为标准, 这个解法有足够高的精确性. 但美中不足的是, 所构造的枢轴量并不是样本 $X_1,\cdots,X_m$ 和 $Y_1,\cdots,Y_n$ 的对称函数. 这样一来, 改变样本 $X_1,\cdots,X_m$ 或 $Y_1,\cdots,Y_n$ 的前后次序, 就得到了不同的区间估计. 而由样本的独立同分布性, 人们很自然地希望, 样本的次序应与问题的解无关. 除了 Scheffe 的精确解法, Welch 在 1938 年给出了 Behrens-Fisher 问题的一个近似解法[1]. Welch 的近似解法较 Scheffe 的精确解法简单, 并且他们给出的区间估计相差也不大, 所以在实际问题中, 通常使用 Welch 的近似解法. 下面对 Welch 的近似解法作一个简单的介绍. 显然,

$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}} \sim N(0,1) \tag{4.25}$$

若将 σ_1^2 和 σ_2^2 分别用其无偏估计 $S_1^2=\dfrac{\sum\limits_{i=1}^{m}(X_i-\overline{X})^2}{m-1}$ 和 $S_2^2=\dfrac{\sum\limits_{i=1}^{n}(Y_i-\overline{Y})^2}{m-1}$ 代替，则得统计量

$$\tilde{t}=\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S}$$

其中 $S^2=\dfrac{S_1^2}{m}+\dfrac{S_2^2}{n}$. 很明显，$\tilde{t}$ 并不服从 t 分布. Welch 通过研究认为，$\tilde{t}$ 的分布近似地与自由度为 r 的 t 分布相同. 关于自由度 r 的值有下面的较为粗糙的解法. 由于 $\tilde{t}$ 是服从标准正态分布 $N(0,1)$ 的随机变量（见(4.25)式）和随机变量 $\sqrt{Z}$ 的商，其中

$$Z=\frac{S^2}{\dfrac{\sigma_1^2}{m}+\dfrac{\sigma_2^2}{n}}$$

所以认为 $\tilde{t}$ 的分布近似地与自由度为 r 的 t 分布相同，相当于认为 Z 的分布近似地与 $\dfrac{W}{r}$ 的分布相同，其中随机变量 $W\sim\chi^2(r)$. 显然，$\mathbf{E}(Z)=\mathbf{E}\left(\dfrac{W}{r}\right)=1$. 由于

$$\mathrm{Var}(Z)=\frac{\dfrac{2\sigma_1^4}{m^2(m-1)}+\dfrac{2\sigma_2^4}{n^2(n-1)}}{\left(\dfrac{\sigma_1^2}{m}+\dfrac{\sigma_2^2}{n}\right)^2}$$

$$\mathrm{Var}\left(\frac{W}{r}\right)=\frac{2}{r}$$

所以

$$r\approx\frac{\left(\dfrac{\sigma_1^2}{m}+\dfrac{\sigma_2^2}{n}\right)^2}{\dfrac{\sigma_1^4}{m^2(m-1)}+\dfrac{\sigma_2^4}{n^2(n-1)}}\tag{4.26}$$

在实用时，(4.26)式中的 σ_1^2 和 σ_2^2 分别用其无偏估计 S_1^2 和 S_2^2 代替，即取

$$r \approx \frac{S^4}{\dfrac{S_1^4}{m^2(m-1)} + \dfrac{S_2^4}{n^2(n-1)}}$$

从而 $\mu_1-\mu_2$ 的置信水平近似为 $1-\alpha$ 的区间估计为

$$\left[\overline{X}-\overline{Y}-S\cdot t_{1-\frac{\alpha}{2}}(r),\quad \overline{X}-\overline{Y}+S\cdot t_{1-\frac{\alpha}{2}}(r)\right]$$

上述方法给出的是置信水平近似为 $1-\alpha$ 的区间估计，下面由信仰推断法给出信仰水平为 $1-\alpha$ 的区间估计.

由例 4.13 知，有函数模型

$$\begin{cases}\mu_1=\overline{X}-\dfrac{1}{\sqrt{m}}\cdot\dfrac{Q_1}{\sqrt{e_2}}\cdot e_1\\[2ex] \sigma_1^2=\dfrac{Q_1^2}{e_2}\\[2ex] \mu_2=\overline{Y}-\dfrac{1}{\sqrt{n}}\cdot\dfrac{Q_2}{\sqrt{f_2}}\cdot f_1\\[2ex] \sigma_2^2=\dfrac{Q_2^2}{f_2}\end{cases}$$

其中 $Q_1^2=\sum_{i=1}^m(X_i-\overline{X})^2, Q_2^2=\sum_{i=1}^n(Y_i-\overline{Y})^2, e_1\sim N(0,1), e_2\sim\chi^2(m-1), f_1\sim N(0,1), f_2\sim\chi^2(n-1), e_1,e_2,f_1$ 和 f_2 相互独立. 由此得

$$\begin{cases}\mu_1=\overline{X}-S_1\cdot t_1/\sqrt{m}\\ \mu_2=\overline{Y}-S_2\cdot t_2/\sqrt{n}\end{cases}$$

其中 $S_1=\dfrac{Q_1}{\sqrt{m-1}}, S_2=\dfrac{Q_2}{\sqrt{n-1}}, t_1\sim t(m-1), t_2\sim t(n-1), t_1$ 和 t_2 相互独立. 从而有

$$\mu_1-\mu_2=\overline{X}-\overline{Y}+\left(S_2\cdot t_2/\sqrt{n}-S_1\cdot t_1/\sqrt{m}\right)$$

其中 $S_2\cdot t_2/\sqrt{n}-S_1\cdot t_1/\sqrt{m}$ 为两个互相独立的 t 分布的线性组合. 我们把它写成如下的形式

$$S_2\cdot t_2/\sqrt{n}-S_1\cdot t_1/\sqrt{m}=r\cdot(\cos\theta\cdot t_2-\sin\theta\cdot t_1)$$

其中 $r=\sqrt{S_1^2/m+S_2^2/n}$，$\cos\theta=\dfrac{S_2/\sqrt{n}}{\sqrt{S_1^2/m+S_2^2/n}}$，$0<\theta<\dfrac{\pi}{2}$. 记 $W=\cos\theta\cdot t_1-\sin\theta\cdot t_2$. 显然，$W$ 的分布仅与 θ 和 t 分布的自由度 $m-1$ 和 $n-1$ 有关. 可以证明：W 的分布关于原点对称. 在 Fisher 和 Yates 编的统计表中可以查得 $w(\theta,m-1,n-1)$，使得

$$P\{-w(\theta,m-1,n-1)\leqslant W\leqslant w(\theta,m-1,n-1)\}=1-\alpha$$

那么 $[\overline{X}-\overline{Y}-r\cdot w(\theta,m-1,n-1),\overline{X}-\overline{Y}+r\cdot w(\theta,m-1,n-1)]$ 就是 $\mu_1-\mu_2$ 的信仰水平为 $1-\alpha$ 的区间估计.

Neyman 通过计算说明，Fisher 给出的信仰水平为 $1-\alpha$ 的区间估计并不是置信水平为 $1-\alpha$ 的区间估计. 所以有的时候信仰水平为 $1-\alpha$ 的区间估计并不是置信水平为 $1-\alpha$ 的区间估计.

Fisher 提出了信仰推断方法. 在他将此方法用于著名的 Behrens-Fisher 问题之后，信仰推断方法受到人们很大的关注. 这说明信仰推断方法是解决区间估计问题的一个有效的方法. 另外，信仰推断方法直观，它容易被实际工作者所接受. 但是，随着研究的深入，人们发现了信仰推断方法的一些内在的问题. 例如，在寻求信仰分布的时候，利用不同的方法，推导出的信仰分布有可能不一样[3],[4]. 信仰分布的这种不确定性是 Fisher 的信仰推断方法至今还没有被人们完全接受的原因之一.

参考文献

1 陈希孺. 数理统计引论. 北京：科学出版社，1981

2 Cox D R and Hinkley D V. Theoretical Statistics. Chapman and Hall, London, 1974

3 Dawid A P and Stone M. The functional-model basis of fiducial inference (with discussion). Ann Statist, 1982, 10: 1054～1074

4 Dawid A P and Wang J L. Fiducial prediction and semi-bayesian inference. Ann Statist, 1993, 21: 1119～1138

习 题 四

4.1 设样本 $X=(X_1,\cdots,X_n)$来自均匀分布总体 $U(0,\theta)$，其中 $\theta>0$. 对给定的 $\alpha(0<\alpha<1)$，构造 θ 的置信水平为 $1-\alpha$ 的置信区间. 如何选取 a 和 b，使得在形如$[b\cdot X_{(n)},a\cdot X_{(n)}]$的 θ 的置信水平为 $1-\alpha$ 的置信区间中，区间的平均长度达到最短？

4.2 设 $X=(X_1,\cdots,X_n)$是来自密度函数为

$$p(x;\theta)=\frac{\theta}{x^2},\quad 0<\theta\leqslant x<\infty$$

的总体的样本. 试求 θ 的置信水平为 $1-\alpha$ 的置信区间. 采用形如$[c\cdot x_{(1)},d\cdot x_{(1)}]$的区间估计. 考虑到区间的平均长度愈短愈好的精度要求，如何选取 c 和 d？

4.3 设样本 $X=(X_1,\cdots,X_n)$来自 Possion 分布总体 $p(\lambda)$，其中 $\lambda>0$.

(1)在样本容量充分大时，试用渐近分布构造 λ 的置信水平近似等于 $1-\alpha$ 的置信限和置信区间；

(2)仿例 4.8，试构造 λ 的精确的置信限和置信区间.

4.4 设 $X=(X_1,\cdots,X_n)$是来自密度函数为

$$p(x;\theta)=\begin{cases}\exp\{-(x-\theta)\}, & x\geqslant\theta\\ 0, & x<\theta\end{cases}$$

的总体的样本. 试基于最小次序统计量 $X_{(1)}=\min\{X_1,\cdots,X_n\}$构造 θ 的置信水平为 $1-\alpha$ 的置信区间.

4.5 设某电子元件的合格品率为 $\theta(0<\theta<1)$. 现对该元件进行测试. 设直到第 $k(k=1,2,\cdots)$次测试才测到合格品. 试构造 θ 的置信水平为 $1-\alpha$ 的置信下限.

4.6 设总体 X 是点集$\{1,2,\cdots,N\}$上的均匀分布，其中 N 为未知参数. 设 $X=(X_1,\cdots,X_n)$是来自 X 的样本. 令 $X_{(n)}=\max\{X_1,\cdots,X_n\}$. 试基于 $X_{(n)}$构造 N 的置信水平为 $1-\alpha$ 的置信上限.

4.7 设 $X=(X_1,\cdots,X_n)$是来自均匀分布 $U(\theta-0.5,\theta+0.5)$的

样本. 试构造 θ 的置信水平为 $1-\alpha$ 的置信区间.

4.8 设 $X=(X_1,\cdots,X_n)$ 是来自均匀分布 $U(\theta_1,\theta_2)$ 的样本.

(1) 试构造 $\theta_2-\theta_1$ 的置信水平为 $1-\alpha$ 的置信区间;

(2) 试构造 $\frac{\theta_1+\theta_2}{2}$ 的置信水平为 $1-\alpha$ 的置信区间.

4.9 设 $X=(X_1,\cdots,X_{10})$ 是来自 Pareto 分布总体的样本. Pareto 分布的密度函数为

$$p(x;\theta)=\begin{cases}\frac{\beta}{\theta}\cdot\left(\frac{\theta}{x}\right)^{\beta+1}, & x\geqslant\theta\\ 0, & \text{其它}\end{cases}$$

其中 $\beta=2$ 已知. 试构造 θ 的置信水平为 $1-\alpha=0.9$ 的 UMA 置信上限.

4.10 令 X 和 Y 分别表示人的脚印长度和身高. 假设 $Y\sim N(X\beta,\sigma^2)$. 现随机抽取 n 个人,它们的脚印长度和身高分别为 $(x_1,y_1;\cdots;x_n,y_n)$.

(1) 试求 β 的 MLE,并构造 β 的置信水平为 $1-\alpha$ 的似然区间估计;

(2) 假设 $n=30$,并经计算得

$$\sum_{i=1}^{n}y_i^2=898\ 044.27,\quad \sum_{i=1}^{n}x_iy_i=130\ 614.6,\quad \sum_{i=1}^{n}x_i^2=19\ 008.24$$

试据此计算 β 的 MLE, β 的置信水平为 $1-\alpha$ 的似然区间估计.

4.11 接例 4.11,试证明, $\hat{\theta}_U^*$ 为 θ 的置信水平为 $1-\alpha$ 的无偏置信上限.

4.12 设样本 $X=(X_1,\cdots,X_n)$ 来自双参数指数分布总体,其密度函数为

$$p(x;\mu,\sigma)=\begin{cases}\frac{1}{\sigma}\cdot\exp\left\{-\frac{x-\mu}{\sigma}\right\}, & x\geqslant\mu\\ 0, & x<\mu\end{cases}$$

(1) 试构造 μ 的置信水平为 $1-\alpha$ 的 UMAU 置信下限;

(2) 试构造 σ 的置信水平为 $1-\alpha$ 的 UMAU 置信区间.

4.13 (Stein 的两阶段抽样方案)设总体 X 服从正态分布 $N(\mu,$

σ^2). 欲构造 μ 的置信水平为 $1-\alpha$ 且区间长度不大于给定正数 L 的置信区间. 关于这个问题,Stein 的两阶段抽样方案如下.

(1)第一阶段抽样:从总体 X 抽取样本 $X_1,\cdots,X_{n_0}$,其中 n_0 是任意预先指定的自然数. 令

$$\overline{X}_0=\frac{1}{n_0}\cdot\sum_{i=1}^{n_0}X_i,\quad S^2=\frac{1}{n_0-1}\cdot\sum_{i=1}^{n_0}(X_i-\overline{X}_0)^2,$$

$$c=\frac{L^2}{4\cdot\left[t_{1-\frac{\alpha}{2}}(n_0-1)\right]^2},\quad n=\max\{n_0,[S^2/c]+1\},$$

($[a]$表示不大于 a 的最大整数);

(2)如果 $n=n_0$,则抽样结束. 构造区间

$$\left[\overline{X}_0-\frac{S}{\sqrt{n_0}}\cdot t_{1-\frac{\alpha}{2}}(n_0-1),\quad \overline{X}_0+\frac{S}{\sqrt{n_0}}\cdot t_{1-\frac{\alpha}{2}}(n_0-1)\right]$$

(3)如果 $n>n_0$,则有第二阶段抽样:从总体 X 继续抽取样本 $X_{n_0+1},\cdots,X_n$. 令

$$\overline{X}_n=\frac{1}{n}\cdot\sum_{i=1}^{n}X_i$$

并构造区间

$$\left[\overline{X}_n-\frac{S}{\sqrt{n}}\cdot t_{1-\frac{\alpha}{2}}(n_0-1),\quad \overline{X}_n+\frac{S}{\sqrt{n}}\cdot t_{1-\frac{\alpha}{2}}(n_0-1)\right]$$

试证明所构造的区间符合要求.

4.14　接习题 4.4,试以 $X_{(1)}$为观察值,θ 为参数构造函数模型,并由此导出 θ 的信仰分布,然后给出 θ 的信仰水平为 $1-\alpha$ 的区间估计. 这个信仰水平为 $1-\alpha$ 的区间估计是不是 θ 的置信水平为 $1-\alpha$ 的区间估计?

4.15　接例 4.14. (1)试求 μ 和 σ^2 的联合信仰分布的密度函数;(2)试证明:在 σ^2 给定的条件下,μ 的条件信仰分布为正态分布 $N\left(\overline{X},\frac{\sigma^2}{n}\right)$;(3)试从区间长度尽可能短的精度要求,分别构造 μ 和 σ^2 的信仰系数为 $1-\alpha$ 的区间估计;(4)试证明:μ 和 σ^2 的信仰系数为 $1-\alpha$ 的区间估计分别是 μ 和 σ^2 的置信水平为 $1-\alpha$ 的区间估计.

4.16 接习题 4.12,试以相互独立的充分统计量$(X_{(1)},\overline{X}-X_{(1)})$为观察值,$(\mu,\sigma)$为参数构造函数模型,并由此导出$(\mu,\sigma)$的信仰分布,以及分别给出$\mu$和$\sigma$的信仰水平为$1-\alpha$的区间估计.这个信仰水平为$1-\alpha$的区间估计是不是置信水平为$1-\alpha$的区间估计?

第五章　统计决策理论与 Bayes 分析

统计决策理论是著名统计学家 A. Wald(1902—1950)在 40 年代建立起来的[1]. 它与经典统计学(如前四章)的差别在于是否涉及后果. 经典统计学着重于在推断上,而不考虑用在何处和效益如何. 而统计决策理论引入损失函数,用来度量效益大小,评价统计推断结果的优劣.

Bayes 分析是英国学者 T. Bayes(1702—1761)首先提出,在本世纪后半叶发展迅速[2]. 它与经典统计学的差别在于是否使用先验信息(经验与历史资料). 经典统计学只用样本信息,而 Bayes 分析把先验信息与样本信息结合起来用于推断之中,形成非决策的 Bayes 分析. 若再使用后果信息,就形成 Bayes 决策分析. 本章将对这些内容的基本部分进行讨论.

§5.1　统计决策问题

§5.1.1　决策问题

我们从双人博弈问题谈起.

例 5.1　设甲乙两人进行一种游戏,甲手中有三张牌,分别标以 $\theta_1,\theta_2,\theta_3$. 乙手中也有三张牌分别标以 a_1,a_2,a_3. 游戏的规则是双方各自独立地出牌,按下表计算甲的得分与乙的失分.

甲\乙	a_1	a_2	a_3
θ_1	3	−2	0
θ_2	1	4	−3
θ_3	−4	−1	2

譬如,甲出 θ_1,乙出 a_1,则甲得 3 分而乙失 3 分. 又如甲出 θ_1,乙出 a_2,则甲失 2 分而乙得 2 分,如此等等. 这张表是甲的得分矩阵,其中负的得分就是失分. 这张表又是乙的失分矩阵,其中负的失分就是得分. 显然,双方都想得高分,但谁也不能控制对方,因此在这种情况下各人该如何"理智"地进行这个游戏便不是一件很简单的事了,假如甲和乙手中的牌更多一些,那此种游戏就更复杂了.

这是一个典型的双人博弈(赌博)问题. 不少实际问题可归结为双人博弈问题. 它是博弈论[7]的研究对象. 如果把上例中的甲方改为自然界或社会,这就是人与自然界(或社会)的博弈问题,这类问题称为决策问题.

例 5.2 某农作物有两个品种:产量高但抗旱能力弱的品种 a_1 和抗旱能力强但产量低的品种 a_2. 在明年雨量不能准确预知的情况下,农民应选播哪个品种可使每亩平均收益最大呢? 这是人与自然界的一局博弈. 在这里一方是人,人是有理智的,他手中有两张牌:a_1 和 a_2. 人手中的牌今后称为行动. 另一方是自然界,它手中的牌是明年雨量,可能有很多张牌,这里为简单起见,以明年 800 mm 雨量为界来区分雨量充足 θ_1 和雨量不充足 θ_2,这样自然界也有两张牌:θ_1 和 θ_2. 由于自然界是无理智的,它的牌今后称为状态或参数. 在这样的格局里,能作决策的仅仅是人. 人为作好决策,可以观察自然界的一些现象,收集和分析自然界的各种信息,也可根据当年种子、肥料、农具等价格定出各种状态下的每亩的收益,然后写出收益矩阵(单位:元):

雨量\人	a_1	a_2
θ_1	1000	300
θ_2	−200	500

自然界虽不会有意与人作对,但也不会任人摆布. 所以人们按此收益矩阵要作决策(选择行动)也并非易事.

这是一个典型的决策问题. 很多实际问题可以归结为这类决策问题. 描述这类决策问题有三个要素:状态集 $\Theta=\{\theta\}$,行动集 $\Delta=\{a\}$和

收益函数 $Q(\theta,a)$，状态集表示自然界(或社会)所有可能状态的全体，行动集表示决策者所有可能采取的行动的全体，收益函数 $Q(\theta,a)$表示在自然界(或社会)处于状态 θ 时，决策者采取行动 a 所获之收益，当 Θ 与 Δ 都是有限集时，$Q(\theta,a)$就成收益矩阵，关于这类问题的决策方法可参见文献[6]. 但很难找到一个较为理想的决策方法. 人们为了更好地作出决策，想方设法从自然界或社会中再去挖掘各种有用的信息. 主要有如下几种信息可供决策使用.

1. 先验信息，人们在过去对自然界(或社会)的各种状态所获得的信息. 这种信息对决策者很起作用. 譬如农民从气象台发布的信息得知，明年雨量偏多，故认为 θ_1(雨量超过 800 mm)发生的概率为 0.6，而 θ_2 发生的概率为 0.4. 这是在 Θ 上的一个分布，常称为先验分布，记为

$$\pi(\theta_1) = 0.6, \quad \pi(\theta_2) = 0.4$$

在此基础上，可算出行动空间 a_1 与 a_2 的平均收益

$$E_\theta[Q(\theta,a_1)] = 1000 \times 0.6 - 200 \times 0.4 = 520(\text{元})$$

$$E_\theta[Q(\theta,a_2)] = 300 \times 0.6 + 500 \times 0.4 = 380(\text{元})$$

从平均收益看，采取行动 a_1 为上策，可见由先验信息形成的先验分布在决策中很起作用.

如果在一个决策问题中还可获得有关 θ 的先验分布，这样的决策问题称为无数据(无样本信息)决策问题. 在文献[6]中有详细讨论，本章不作研究.

2. 样本信息，从与自然界(或社会)的状态 θ 有关的环境中去抽样，从获得的样本中了解当今状态 θ 的最新信息，这里的关键是要确定一个概率分布，使 θ 恰好是其未知参数.

如果在一个决策问题中还利用样本信息，这种的问题称为统计决策问题. 如果在一个统计决策问题中还利用先验信息，这样的问题称为 Bayes 决策问题. 本章先讲述统计决策问题，然后再讲 Bayes 决策问题. 为了更好地认识统计决策问题，我们先看下面一个例子.

例 5.3 某工厂的产品每 100 件装成一箱运交顾客，在向顾客交货前面临如下两个行动.

a_1：一箱中逐一检查；

a_2:一箱中都不检查.

若工厂选择行动 a_1,则可保证交货时每件产品都是合格品. 但因每件产品的检查费为 0.8 元,为此工厂要支付检查费 80 元/箱. 若工厂选择行动 a_2,工厂可免付每箱检查费 80 元,但顾客发现不合格品时,按合同不仅允许更换,而且每件还要支付 12.5 元的赔偿费. 按此算得:

若一箱中不合格品不超过 6 件,赔偿费不超过 75 元,那选择行动 a_2 比 a_1 有利;

若一箱中不合格品不少于 7 件,赔偿费不低于 87.5 元,那选择行动 a_1 比 a_2 有利.

为了得知一箱中的不合格品率 θ,工厂决定先在每箱中抽取两件进行检查,X 为其不合格品的件数,根据 X 的取值(可能取 0,1,2 三种)再选择行动 a_1 或 a_2. 这时工厂的支付函数可算得

$$W(\theta,a)=\begin{cases}80, & a=a_1\\ 1.6+1250\theta, & a=a_2\end{cases} \tag{5.1}$$

如何依据抽样结果和支付函数作出决策,使工厂的支付费用最少呢?这是一个典型的统计决策问题.

§ 5.1.2 统计决策问题的三个基本要素

构成一个统计决策问题有如下三个基本要素.

1. 可控参数统计结构$(\mathscr{X},\mathscr{B},\{p_\theta:\theta\in\Theta\})$,其中参数空间 Θ 中每个元素就是自然界或社会可能处的状态. 从概率密度函数 $p_\theta(x)$ 抽样,获得样本 $X=(X_1,\cdots,X_n)$,就是为了从此样本获取有关 θ 的最新信息,以便更好地作出决策. 在具体问题中,当实际背景了解得很清楚时,选择和确定这样的结构是不难做到的. 譬如,在例 5.3 中,选用二项分布族形成的统计结构是很自然的事.

2. 行动空间$(\Delta,\mathscr{B}_\Delta)$,其中 $\Delta=\{a\}$ 是为解决某统计决策问题时,人们对自然界(或社会)可能作出的一切行动的全体. Δ 中的每个元素表示一个行动. $\mathscr{B}_\Delta$ 是 Δ 上的某个 σ 代数,这是为以后扩充概念而假设的. 譬如,在例 5.3,$\Delta=\{a_1,a_2\}$ 中只含两个行动,其中 a_1 表示一箱中的产品每件都检查,a_2 表示一箱中的产品每件都不检查. 这两类行动决

策问题在实际中常会遇到.

一个行动空间 Δ 至少应含有两个行动，假如 Δ 中只有一个行动，那人们就无需选择、从而也形成不了一个统计决策问题.

3. 损失函数 $L(\theta,a)$，它是定义在 $\Theta\times\Delta$ 上的二元函数. 它表示当自然界(或社会)处于状态 θ 时，而人们采取行动 a 对人们引起的(经济)损失. 损失函数 $L(\theta,a)$ 是把决策与经济效益联系在一起的桥梁. 有了它决策才能进入定量分析阶段. 它要根据实际情况确定. 是三个要素中最重要的要素. 当 Θ 和 Δ 都是有限集时，损失函数就成了损失矩阵.

这里的损失函数是从收益函数、亏损函数、支付函数、成本函数等概念引申出的更一般更有效的概念. 譬如，某商店一个月的经营收益为 -1000 元即亏 1000 元. 这是对成本而言. 在这里不称其为损失，而称其为亏损. 我们讲的损失是指"该赚而没有赚到的钱"，或"不该亏而亏损的钱"或"不该支付而支付的钱". 譬如，一商店本可赚 2000 元，由于决策失误而亏了 1000 元，那我们说该商店损失了 3000 元，又如，一次加班只需 5 位工人就可完成，可去了 7 位工人，工作虽是完成了，可工厂多支付了 2 位工人的加班费就是损失. 用这种观点去认识损失对提高决策意识是很有好处的.

从收益函数很容易获得损失函数，先看下面的例子.

例 5.4　某公司购进某种货物可分为大批、中批和小批三种行动(依次记为 a_1,a_2,a_3). 未来市场需求量可分为高、中和低三种状态(依次记为 $\theta_1,\theta_2,\theta_3$). 这三种行动在不同市场状态下获得的利润如下(单位：千元)

$$Q=\begin{array}{c} \begin{array}{ccc} a_1 & a_2 & a_3 \end{array} \\ \left(\begin{array}{ccc} 10 & 5 & 2 \\ 3 & 4 & 2 \\ -2.7 & -0.8 & 1 \end{array}\right) \begin{array}{c} \theta_1 \\ \theta_2 \\ \theta_3 \end{array} \end{array}$$

这是一个收益矩阵. 现我们按损失的含义把它改写为损失矩阵.

当市场处于状态 $\theta_1(\theta=\theta_1)$ 时，从 Q 的第一行可以看出，这时的最优行动是 a_1，可收益 10 000 元. 倘若经理采取行动 a_2，这时收益仅为 5000 元. 与最优行动 a_1 相比，要少得 10 000－5000＝5000 元，这 5000

元为公司"该赚而未赚到的钱",是 $\theta=\theta_1, a=a_2$ 时经理的损失值. 即 $L(\theta_1,a_2)=5000$元,类似地,$L(\theta_1,a_3)=10\ 000-2000=8000$ 元. 而 $L(\theta_1,a_1)=0$(即无损失). 这是因为当 θ_1 发生时,经理所采取的是最优行动 a_1. 用同样的方法可算出 θ 与 a 在不同情况下的一切损失值. 假如把这些损失按原来次序排成一个矩阵(单位:千元)

$$
\begin{array}{cc} & \begin{array}{ccc} a_1 & a_2 & a_3 \end{array} \\
L = & \begin{pmatrix} 0 & 5 & 8 \\ 1 & 0 & 2 \\ 3.7 & 1.8 & 0 \end{pmatrix} \begin{array}{c} \theta_1 \\ \theta_2 \\ \theta_3 \end{array}
\end{array}
$$

这个矩阵称为经理的损失矩阵. 从上述转换过程中可以看出. 损失值是指经理该赚而未赚到的钱. 经理在做决策时,要尽量避免大损失,追求无损失. 只有经理的决策为最优时损失才会为零.

一般说来,当收益函数为 $Q(\theta,a)$时,自然界(或社会)处于 θ 时的最大收益为$\max\limits_{a\in\varDelta}Q(\theta,a)$,而人们采取行动 a 的损失值为

$$L(\theta,a)=\max_{a\in\varDelta}Q(\theta,a)-Q(\theta,a) \tag{5.2}$$

类似地,当决策用支付函数 $W(\theta,a)$时(如例 5.3 所示),它是越小越好. 自然界(或社会)处于 θ 时的最小支付为$\min\limits_{a\in\varDelta}W(\theta,a)$,而人们采取行动 a 的损失值为

$$L(\theta,a)=W(\theta,a)-\min_{a\in\varDelta}W(\theta,a) \tag{5.3}$$

例 5.5　某公司购进一批货物投放市场,若购进数量 a 低于市场需求量 θ(即 $a\leqslant\theta$),每吨可赚 15 万元. 若购进数量 a 超过市场需求量 θ(即 $a>\theta$),超过部分每吨反要亏 35 万元. 由此可写出其收益函数

$$Q(\theta,a)=\begin{cases}15a, & a\leqslant\theta \\ 15\theta-35(a-\theta), & a>\theta\end{cases}$$

容易看到,在购销过程无损耗时,购进量 a 等于市场需求量 θ 时,收益达到最大,此时收益为 $Q(\theta,\theta)=15\theta$. 再根据(5.2)式,立即可写出其损失函数为

$$L(\theta,a)=\begin{cases}15(\theta-a), & a\leqslant\theta \\ 35(a-\theta), & a>\theta\end{cases}$$

该损失函数表明：当购进数量 a 低于市场需求量 θ 时，有 $15(\theta-a)$ 万元的钱是该赚而没有赚到. 这是损失，当购进数量 a 高于市场需求量 θ 时，有 $35(a-\theta)$ 万元的钱是不该亏的而亏了，这也是损失.

在统计决策问题的三个基本要素中最重要的是损失函数. 它把决策行动与后果联系起来了. 有了损失函数就等于把后果信息引入到统计问题中来了.

§5.1.3　常用的损失函数

损失函数 $L(\theta,a)$ 是要根据实际问题而定的. 但不论在什么场合，今后总要求损失函数是非负的，即

$$L(\theta,a) \geqslant 0, \qquad \forall\, \theta \in \Theta, a \in \Delta \tag{5.4}$$

这等价于把最小的损失定为零，这不仅不失损失函数的含义，而且对今后研究带来方便.

一般总认为：行动 a 离状态 θ 愈远而引起的损失愈大. 所以损失函数 $L(\theta,a)$ 应是距离 $|a-\theta|$ 的非降函数. 常取

$$L(\theta,a) = \lambda(\theta) \cdot g(|a-\theta|) \tag{5.5}$$

其中 $\lambda(\theta)>0$，且有限，它反映决策中，由于 θ 的不同，即使同一个偏差 $|a-\theta|$ 造成的危害性常不一样. 而 $g(t)$ 是 t 的非降函数. 最常见的形式是

$$L(\theta,a) = \lambda(\theta)\,|a-\theta|^k$$

其中 k 常取非负整数，最常用的形式是如下几种.

1. 平方损失函数

$$L(\theta,a) = (a-\theta)^2 \tag{5.6}$$

或加权平方损失函数

$$L(\theta,a) = \lambda(\theta)(a-\theta)^2 \tag{5.7}$$

这是在统计决策问题中用得最多的损失函数. 人们之所以喜爱用这个损失函数部分理由是把 $g(t)$ 在零点展开成 Taylor 级数时，二次函数是一种很好的近似，另一部分的理由是它与最小二乘法的形式相似，将来计算方便，又有一些很好性质. 它的失真之处在于不是上有界，致使损失随着偏差 $|a-\theta|$ 增大而按平方律增大，直至无穷. 这会使人很担忧.

2. 线性损失函数

$$L(\theta,a)=\begin{cases}K_0(\theta-a), & a\leqslant\theta\\ K_1(a-\theta), & a>\theta\end{cases} \tag{5.8}$$

其中 K_0 和 K_1 是两个常数.它们的选择常反映行动 a 低于状态 θ 和高于状态 θ 的相对重要性.例 5.5.中的损失函数就是线性的.

当 $K_0=K_1$ 时,就得绝对损失函数

$$L(\theta,a)=|a-\theta| \tag{5.9}$$

若 K_0,K_1 分别是 θ 的函数,则称加权线性损失函数

$$L(\theta,a)=\begin{cases}K_0(\theta)(\theta-a), & a\leqslant\theta\\ K_1(\theta)(a-\theta), & a>\theta\end{cases} \tag{5.10}$$

3. 0-1 损失函数

$$L(\theta,a)=\begin{cases}0, & |a-\theta|\leqslant\varepsilon\\ 1, & |a-\theta|>\varepsilon\end{cases} \tag{5.11}$$

这里的 ε 是正数,这种损失函数常在两行动决策问题中使用.这里的 0 与 1 与其说是损失大小,还不如说是有无损失的示意.特别当有损失时,其损失很少与真实损失近似.更现实的损失可能是

$$L(\theta,a)=\begin{cases}0, & |a-\theta|\leqslant\varepsilon\\ k, & |a-\theta|>\varepsilon\end{cases} \tag{5.12}$$

或

$$L(\theta,a)=\begin{cases}0, & |a-\theta|\leqslant\varepsilon\\ k(\theta), & |a-\theta|>\varepsilon\end{cases} \tag{5.13}$$

4. 多元二次损失函数,当 $\boldsymbol{\theta}$ 和 $\boldsymbol{a}$ 均为多维向量时,可取如下二次型作为损失函数

$$L(\boldsymbol{\theta},\boldsymbol{a})=(\boldsymbol{a}-\boldsymbol{\theta})'\boldsymbol{A}(\boldsymbol{a}-\boldsymbol{\theta}) \tag{5.14}$$

其中 $\boldsymbol{\theta}'=(\theta_1,\cdots,\theta_p)$,$\boldsymbol{a}'=(a_1,\cdots,a_p)$,$\boldsymbol{A}$ 为 $p\times p$ 阶正定阵.当 $\boldsymbol{A}$ 为对角阵时,即 $\boldsymbol{A}=\text{diag}(w_1,\cdots,w_p)$,则此 p 元损失函数为

$$L(\boldsymbol{\theta},\boldsymbol{a})=\sum_{i=1}^{p}w_i(a_i-\theta_i)^2 \tag{5.15}$$

其中诸 w_i 可看作为各参数重要性的加权.

上述损失函数大多用于状态 θ 和行动 a 都是连续变量场合.实际

中常会遇到行动空间 Δ 只含有限个行动的问题，二行动线性决策问题就是其中最简单的一类. 下面以例子形式加以叙述.

例 5.6　设某一统计决策问题只有两个行动可供选择，即 $\Delta=\{a_1, a_2\}$，而状态 θ 可以是离散的，也可以是连续的. 在这种场合，由于行动只有两个，很容易获得每个行动下的收益函数或支付函数. 假如这些函数都是 θ 的线性函数，譬如，其支付函数为

$$W(\theta,a)=\begin{cases}b_1+m_1\theta, & a=a_1\\ b_2+m_2\theta, & a=a_2\end{cases} \tag{5.16}$$

其中 $m_1<m_2$，b_1，b_2 皆为已知函数. 这类问题称为二行动线性决策问题.

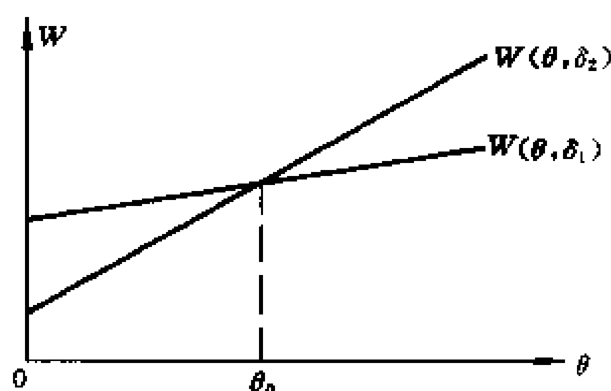


图 5.1　线性支付函数（$m_1<m_2$）

在这类问题中两个线性函数常有一个交点（见图 5.1），此交点容易算得 $\theta_0=(b_1-b_2)/(m_2-m_1)$. 由损失函数的定义（5.3）容易算得 a_1 和 a_2 时的损失函数值，当 $\theta\leqslant\theta_0$ 时，

$$\begin{aligned}L(\theta,a_1)&=W(\theta,a_1)-\min_{a\in\Delta}W(\theta,a)\\&=b_1+m_1\theta-(b_2+m_2\theta)\\&=b_1-b_2+(m_1-m_2)\theta\\L(\theta,a_2)&=W(\theta,a_2)-\min_{a\in\Delta}W(\theta,a)=0\end{aligned}$$

类似地，当 $\theta>\theta_0$ 时，有

$$L(\theta,a_1)=0$$
$$L(\theta,a_2)=(b_2-b_1)+(m_2-m_1)\theta$$

由此可得二行动线性决策问题的损失函数

$$\begin{aligned} L(\theta,a_1) &= \begin{cases} b_1 - b_2 + (m_1 - m_2)\theta, & \theta \leqslant \theta_0 \\ 0, & \theta > \theta_0 \end{cases} \\ L(\theta,a_2) &= \begin{cases} 0, & \theta \leqslant \theta_0 \\ b_2 - b_1 + (m_2 - m_1)\theta, & \theta > \theta_0 \end{cases} \end{aligned} \tag{5.17}$$

前述的例 5.3 就是二行动线性决策问题，在那里支付函数如(5.1)所示，相应的损失函数可算得

$$L(\theta,a_1) = \begin{cases} 78.4 - 1250\theta, & \theta \leqslant \theta_0 \\ 0, & \theta > \theta_0 \end{cases}$$

$$L(\theta,a_2) = \begin{cases} 0, & \theta \leqslant \theta_0 \\ -78.4 + 1250\theta, & \theta > \theta_0 \end{cases}$$

其中 $\theta_0 = 0.062\ 72$. 对线性收益函数 $Q(\theta,a)$亦可类似获得损失函数.

§ 5.2 决策函数和风险函数

当给出一个统计结构$(\mathscr{X},\mathscr{B},\{p_\theta:\theta\in\Theta\})$，一个行动空间$(\Delta,\mathscr{B}_\Delta)$和一个损失函数 $L(\theta,a)$等三要素后，就确定了一个统计决策问题. 为了在一个统计决策问题中选择行动进行决策，我们需要决策函数与风险函数等概念.

§ 5.2.1 决策函数

定义 5.1 在一个统计决策问题中，从样本空间$(\mathscr{X},\mathscr{B})$到行动空间$(\Delta,\mathscr{B}_\Delta)$的可测映照 $\delta(x)$称为(非随机化)决策函数.

这里要求 $\delta(x)$可测是使有关 $\delta(x)$的事件有概率可言. 而"非随机化"是指样本 X 一经确定，决策者所采取的行动 $\delta(x)$也唯一确定. 再无任何随机性可言. 而当样本确定，$\delta(x)$尚不能唯一确定，还要按一定的概率分布在 Δ 中抽样才能确定，这样的决策函数将称为随机化决策函数. 它将在后面详细叙述.

从上述定义可见，当行动空间 Δ 是某个实数集时，上述决策函数

就是统计量，可决策函数还允许其值不是实数而是某个行动. 在例5.3中由 $\mathscr{X}=\{0,1,2\}$ 到 $\Delta=\{a_1,a_2\}$ 的一切可能的变换都是决策函数，其中每个决策函数的取值不是实数，而是特定的行动 a_1（全数检查）或 a_2（一个也不检查）.

§5.2.2 风险函数

定义5.2 设 $\delta(X)$ 是一个统计决策问题中的决策函数，那么损失函数 $L(\theta,\delta(X))$ 关于样本 X 的分布 $p_\theta(x)$ 的数学期望

$$R(\theta,\delta)=\mathrm{E}_{X|\theta}[L(\theta,\delta(X))]=\int_{\mathscr{X}}L(\theta,\delta(x))p_\theta(x)\mathrm{d}\mu \quad (5.18)$$

称为决策函数 $\delta(X)$ 的风险函数. 有时简称风险. 其中 μ 是控制测度，常取 Lebesgue 测度或计数测度.

从上述定义可见，风险就是平均损失，平均损失是愈小愈好. 它是用来度量决策函数好坏的一把尺子. 当决策函数 $\delta(X)$ 给定时，风险函数仍是 θ 的函数. 比较两个决策函数 $\delta_1(X)$ 和 $\delta_2(X)$ 的好坏就要看其风险函数 $R(\theta,\delta_1)$ 和 $R(\theta,\delta_2)$ 的大小. 比较两个函数的大小是困难的. 少数场合（见图5.2(*a*)）可以立即看出谁好谁坏. 多数场合（见图5.2(*b*)）是不易区分的.

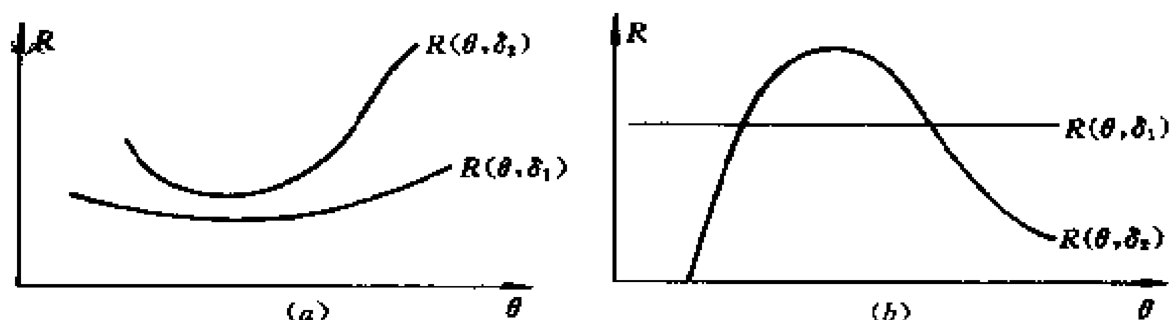


图5.2 两个风险函数的比较

几种可比较风险函数大小的场合可由下面的定义给出.

定义5.3 设 $\delta_1(X)$ 和 $\delta_2(X)$ 是统计决策问题中的两个决策函数. 假如其风险函数间有

$$R(\theta,\delta_1) \leqslant R(\theta,\delta_2), \quad \forall\ \theta \in \Theta$$

且存在一些 θ 使不等式成立. 则称决策函数 $\delta_1(X)$一致优于 $\delta_2(X)$. 假如其风险函数间有

$$R(\theta,\delta_1) = R(\theta,\delta_2), \quad \forall\ \theta \in \Theta$$

则称决策函数 $\delta_1(X)$与 $\delta_2(X)$等价.

定义 5.4 设 $\mathscr{D}=\{\delta(X)\}$是一切可能定义在$(\mathscr{X},\mathscr{B})$上而在$(\Delta,\mathscr{B}_\Delta)$上取值的决策函数的全体. 假如在决策函数类 $\mathscr{D}$ 中存在这样一个决策函数 $\delta^*(X)$,使得对任一个 $\delta(X)\in\mathscr{D}$ 皆有

$$R(\theta,\delta^*) \leqslant R(\theta,\delta), \quad \forall\ \theta \in \Theta$$

则称 $\delta^*(X)$为(该决策函数类 $\mathscr{D}$ 的)一致最小风险决策函数,或称一致最优决策函数.

上述两个定义都是对某一个给定损失函数而言的. 当损失函数改变了,结论可能不成立了. 在定义 5.4 中,还要对某一决策函数类而言的. 当决策函数类改变了,一致最优性可能就不具备了.

例 5.7 设 x_1 和 x_2 是从下列分布获得的两个观察值

$$P_\theta(X = \theta - 1) = P_\theta(X = \theta + 1) = 0.5, \quad \theta \in \Theta = \mathbf{R}$$

现把它放在统计决策问题的框架下来研究 θ 的估计问题,为此取行动空间 $\Delta=\boldsymbol{R}$,损失函数为

$$L(\theta,a) = 1 - I_\theta(a)$$

其中 $I_\theta(a)$为示性函数,当 $a=\theta$ 时它为 1,否则为 0,大家知道,从样本空间 $\mathscr{X}=\{(x_1,x_2)\}$到行动空间 Δ 上的决策函数有许多,现考察其中三个.

(1) $\delta_1(x_1,x_2)=(x_1+x_2)/2$,其风险函数为

$$R(\theta,\delta_1) = 1 - P_\theta(\delta_1 = \theta) = 1 - P_\theta(x_1 \neq x_2) = 0.5, \quad \forall\ \theta \in \Theta$$

(2) $\delta_2(x_1,x_2)=x_1-1$,其风险函数为

$$\begin{aligned} R(\theta,\delta_2) &= 1 - P_\theta(\delta_2 = \theta) \\ &= 1 - P_\theta(x_1 = \theta + 1) = 0.5, \quad \forall\ \theta \in \Theta \end{aligned}$$

(3) $\delta_3(x_1,x_2)=\begin{cases}(x_1+x_2)/2, & x_1\neq x_2, \\ x_1-1, & x_1=x_2,\end{cases}$ 其风险函数为

$$R(\theta,\delta_3)=1-P_\theta(\delta_3=\theta)$$
$$=1-P_\theta(x_1\neq x_2 \text{ 或 } x_1=\theta+1)=0.25,\quad \forall\,\theta\in\Theta$$

假如只限于考察这三个决策函数组成的类 $\mathscr{D}=\{\delta_1,\delta_2,\delta_3\}$，那么 δ_3 是此决策函数类中一致最优决策函数，当决策函数类扩大或损失函数改变时，δ_3 的最优性可能会消失.

例 5.8 在例 5.3 中我们讨论了一个二行动线性统计决策问题. 在那里有两个行动可供选择，一是每箱 100 件产品逐一检查(a_1)；另一是每箱中一件都不检查(a_2)，工厂决定：先从每箱中任取两件检查，然后根据检查结果再作决策，这里抽样检查结果为 0，1，2 等不合格品件数. 由此可见，从 $\mathscr{X}=\{0,1,2\}$ 到 $\Delta=\{a_1,a_2\}$ 上的任一变换都是该统计决策问题的决策函数. 此种决策函数共有 8 个，它们是

x	0	1	2
$\delta_1(x)$	a_1	a_1	a_1
$\delta_2(x)$	a_1	a_1	a_2
$\delta_3(x)$	a_1	a_2	a_1
$\delta_4(x)$	a_1	a_2	a_2
$\delta_5(x)$	a_2	a_1	a_1
$\delta_6(x)$	a_2	a_1	a_2
$\delta_7(x)$	a_2	a_2	a_1
$\delta_8(x)$	a_2	a_2	a_2

譬如，$\delta_5(x)$仅在 $x=0$(抽两件产品全是合格品)时采取行动 a_2(一件都不查)，而 x 在为 1 或 2 时采取行动 a_1(逐一检查)，或者写为

$$\delta_5(x)=\begin{cases}a_1, & x=1,2\\ a_2, & x=0\end{cases}$$

在抽检两件产品后，工厂为每箱的支付函数 $W(\theta,a)$已在例 5.3 中指出，在例 5.6 中还把它改写为如下损失函数

$$L(\theta,a_1)=\begin{cases}78.4-1250\theta, & \theta\leqslant\theta_0\\ 0, & \theta>\theta_0\end{cases}$$

$$L(\theta,a_2)=\begin{cases}0, & \theta\leqslant\theta_0\\ -78.4+1250\theta, & \theta>\theta_0\end{cases}$$

其中 $\theta_0=0.062\ 72$. 另外若设 X 为抽取两件产品中的不合格品的件数，则 X 服从二项分布 $b(2,\theta)$，即

$$P_\theta(X=x)=\binom{2}{x}\theta^x(1-\theta)^{2-x},\quad x=0,1,2$$

由此可算得 8 个决策函数的风险函数. 譬如 $a_5(x)$ 的风险函数为

$$\begin{aligned}R(\theta,\delta_5(x))&=\mathrm{E}_{x|\theta}L(\theta,\delta_5(X))\\&=L(\theta,\delta_5(0))P_\theta(X=0)+L(\theta,\delta_5(1))P_\theta(X=1)+\\&\quad L(\theta,\delta_5(2))P_\theta(X=2)\\&=L(\theta,a_2)(1-\theta)^2+L(\theta,a_1)\cdot 2\theta(1-\theta)+L(\theta,a_1)\theta^2\end{aligned}$$

把损失函数代入，当 $\theta\leqslant\theta_0$ 时

$$R(\theta,\delta_5(x))=(78.4-1250\theta)[1-(1-\theta)^2]$$

当 $\theta>\theta_0$ 时

$$R(\theta,\delta_5(x))=(-78.4+1250\theta)(1-\theta)^2$$

类似地可写出其它几个决策函数的风险函数. 它们是

$$R(\theta,\delta_1(x))=\begin{cases}78.4-1250\theta, & \theta\leqslant\theta_0\\0, & \theta>\theta_0\end{cases}$$

$$R(\theta,\delta_2(x))=\begin{cases}(78.4-1250\theta)(1-\theta^2), & \theta\leqslant\theta_0\\(-78.4+1250\theta)\theta^2, & \theta>\theta_0\end{cases}$$

$$R(\theta,\delta_3(x))=\begin{cases}(78.4-1250\theta)[1-2\theta(1-\theta)], & \theta\leqslant\theta_0\\(-78.4+1250\theta)\cdot 2\theta(1-\theta), & \theta>\theta_0\end{cases}$$

$$R(\theta,\delta_4(x))=\begin{cases}(78.4-1250\theta)(1-\theta)^2, & \theta\leqslant\theta_0\\(-78.4+1250\theta)[1-(1-\theta)^2], & \theta>\theta_0\end{cases}$$

$$R(\theta,\delta_6(x))=\begin{cases}(78.4-1250\theta)\cdot 2\theta(1-\theta), & \theta\leqslant\theta_0\\(-78.4+1250\theta)[1-2\theta(1-\theta)], & \theta>\theta_0\end{cases}$$

$$R(\theta,\delta_7(x))=\begin{cases}(78.4-1250\theta)\theta^2, & \theta\leqslant\theta_0\\(-78.4+1250\theta)(1-\theta^2), & \theta>\theta_0\end{cases}$$

$$R(\theta,\delta_8(x))=\begin{cases}0, & \theta\leqslant\theta_0\\(-78.4+1250\theta), & \theta>\theta_0\end{cases}$$

这些风险函数在 $\theta=0(0.02)0.12$ 的值已列在表 5.1 中.

表 5.1 8 个风险函数值($\theta=0(0.02)0.12$)(单位:元)

$R(\theta,\delta(x))$	0	0.02	0.04	0.06	0.08	0.10	0.12
$R(\theta,\delta_1(x))$	78.4	53.40	28.40	3.40	0	0	0
$R(\theta,\delta_2(x))$	78.4	53.38	28.35	3.39	0.14	0.47	1.03
$R(\theta,\delta_3(x))$	78.4	51.31	26.22	3.02	3.14	8.39	15.12
$R(\theta,\delta_4(x))$	78.4	51.29	26.17	3.00	3.32	8.85	16.15
$R(\theta,\delta_5(x))$	0	2.11	2.23	0.40	18.68	37.75	55.45
$R(\theta,\delta_6(x))$	0	2.09	2.18	0.38	18.42	38.21	56.48
$R(\theta,\delta_7(x))$	0	0.02	0.05	0.01	21.46	46.13	70.57
$R(\theta,\delta_8(x))$	0	0	0	0	21.60	46.60	71.60

从表 5.1 上可以看出:

1. 若该厂的不合格品率 θ 在 0 到 0.12 之间,那在此 8 个决策函数组成的类中不存在一致最优决策函数.

2. 若该厂的不合格品率 θ 在 0 到 0.06 之间,那在此决策函数类中 $\delta_8(x)=a_2$(一件也不检查)是一致最优决策函数.

3. 若该厂的不合格品率 θ 在 0.07 到 0.12 之间,那在此决策函数类中 $\delta_1(x)=a_1$(逐件检查)是一致最优决策函数.

4. 若该厂的不合格品率 θ 在 0.04 到 0.08 之间,用"先抽样,后决策"的方法可以使平均损失大为减少.但仍选不出一致最优决策函数.

§5.2.3 经典统计推断三种基本形式的再描述

经典统计推断的三种基本形式:点估计,区间估计和假设检验都可以看作一种特殊的统计决策问题.这里先从统计决策观点看点估计问题.

1. 点估计问题.

设有一个可控参数结构$(\mathscr{X},\mathscr{B},\{p_\theta:\theta\in\Theta\})$,又设 $X=(X_1,\cdots,X_n)$是其一个样本.在寻求参数 θ 的点估计问题中,我们把行动空间 Δ 取为参数空间 Θ,即 $\Delta\equiv\Theta$. 估计量 $\hat\theta=\hat\theta(X_1,\cdots,X_n)$就是从样本空间

$\mathscr{X}$ 到行动空间 Δ 上的一个决策函数. 损失函数 $L(\theta,\hat{\theta})$ 可解释为当 θ 为真参数时，而用 $\hat{\theta}$ 去估计它们所引起的损失. 这样一来，点估计问题就是一个特殊的统计决策问题.

假如损失函数为平方损失函数

$$L(\theta-\hat{\theta})=(\hat{\theta}-\theta)^2$$

那么风险函数 $R(\theta,\hat{\theta})=\mathrm{E}_{X|\theta}(\hat{\theta}-\theta)^2$ 就是均方误差，这时最小均方估计就是 θ 的一切决策函数类 $\mathscr{D}$ 中的一致最优决策函数，在第二章中曾指出此种估计并不存在. 假如把决策函数类限于 θ 的无偏估计类 $\mathscr{D}_1$ 中，那么其风险函数就是方差，这时 θ 的一致最小方差无偏估计(UMVUE)就是 $\mathscr{D}_1$ 中的一致最优决策函数. 从决策理论看，最小方差不是一个估计为"最优"的唯一标准. 改变损失函数后就可得另一种意义下的最优估计. 譬如，Pitman 考虑寻找这样一个估计 $\theta^*(X_1,\cdots,X_n)$，使得 $|\theta^*-\theta|\leqslant c$ 的概率为最大，其中 c 为某事先给定的正数. 这种估计问题也可看作一个特殊的统计决策问题，这只要取 0-1 损失函数

$$L(\theta,\theta^*)=\begin{cases}0, & |\theta^*-\theta|\leqslant c\\ 1, & |\theta^*-\theta|>c\end{cases}$$

于是寻找 Pitman 意义下的最优估计，就是在 0-1 损失函数下寻找最小风险决策函数问题.

2. 区间估计问题

设有一个可控参数结构$(\mathscr{X},\mathscr{B},\{p_\theta:\theta\in\Theta\})$，又设 $X=(X_1,\cdots,X_n)$是其中一个样本. 在寻求参数 θ 的区间估计问题中，我们可把行动空间 Δ 取为某个给定的区间类，如

Δ_1：直线上所有的有界区间类

Δ_2：直线上长度不超过给定值 c 的区间类

这时 Δ 中每个行动就是一个区间，决策函数就是定义在 $\mathscr{X}$ 上而在 Δ 中取值的区间函数.

$$\delta(x)=(d_1(x),d_2(x))$$

其中 $d_1(x)$和 $d_2(x)$为区间的两个端点. 若取如下损失函数

$$L(\theta,\delta(x))=m_1(d_2-d_1)+m_2[1-I_{(d_1,d_2)}(\theta)]$$

其中 m_1 和 m_2 为某两个给定正常数. 第一项表示这个区间 (d_1,d_2) 长短引起的损失,第二项表达了当 θ 不属于区间 (d_1,d_2) 时而引起的损失,这时其风险函数为

$$R(\theta,\delta(X)) = m_1 \mathrm{E}_{X|\theta}(d_2 - d_1) + m_2 P(\theta \notin (d_1,d_2))$$

其中第一项与平均长度成比例,第二项与区间不包含真值的概率成比例.

假如 $X_1,\cdots,X_n$ 是来自正态总体 $N(\theta,\sigma^2)$ 的一个样本,那么在给定置信水平 $1-\alpha(0<\alpha<1)$ 后,可用 t 统计量获得 θ 的置信区间

$$(d_1(x),d_2(x)) = \left(\bar{x} - \frac{s}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1), \bar{x} + \frac{s}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)\right)$$

其中 $\bar{x}$ 为样本均值, $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$, $t_{\frac{\alpha}{2}}(n-1)$ 为是自由度为 $n-1$ 的 t 分布的 $\alpha/2$ 上侧分位数. 于是其平均长度为

$$\mathrm{E}(d_2 - d_1) = \frac{2}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)\cdot \mathrm{E}(S)$$

$$= \frac{2\sqrt{2}}{\sqrt{n(n-1)}}\sigma t_{\frac{\alpha}{2}}(n-1)\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}$$

而该区间不包含真值 θ 的概率为

$$P[\theta \notin (d_1,d_2)] = \alpha$$

所以最后的风险函数为

$$R(\theta,(d_1,d_2)) = \frac{2\sqrt{2}}{\sqrt{n(n-1)}}m_1\sigma t_{\frac{\alpha}{2}}(n-1)\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} + m_2\alpha$$

在这样的统计决策问题中要寻找一致最优决策函数也是办不到的,即使在 $m_1=0$ 或 $m_2=0$ 场合. 寻找一致最优(平均长度最短或包含错误值的概率最小)决策函数也只在特定的区间类中才能实现(见第四章).

3. 假设检验问题

设有一个可控参数结构 $(\mathscr{X},\mathscr{B},\{P_\theta: \theta\in\Theta\})$, 又设 Θ_0 和 Θ_1 是 Θ 的两个不相交的非空子集. 又设 $X=(X_1,\cdots,X_n)$ 是其一个样本,在原

假设 Θ_0 对备择假设 Θ_1 的检验问题中，我们常把行动空间 Δ 取作为仅有两个行动组成的集合，有时就令

$$\Delta = \{0,1\}$$

其中“0”表示接受原假设 Θ_0，“1”表示拒绝原假设（接受备择假设）这时损失函数 $L(\theta,a)$ 可解释为当 θ 为真时，采取行动 a（取 0 或 1）所引起的损失. 这样一来，原假设 Θ_0 对备择假设 Θ_1 的检验问题就可看作一个统计决策问题.

在这个决策问题中，决策函数 $\delta(X)$ 就是样本空间 $\mathscr{X}$ 到行动空间 Δ 的一可测变换. 若记

$$W = \{(x_1,\cdots,x_n) : \delta(x_1,\cdots,x_n) = 1\} \subset \mathscr{X}$$

则任一决策函数可表示为

$$\delta(x_1,\cdots,x_n) = \begin{cases} 1, & (x_1,\cdots,x_n) \in W \\ 0, & (x_1,\cdots,x_n) \notin W \end{cases}$$

这表明：任一个决策函数是拒绝域的示性函数.

假如损失函数取如下形式

$$L(\theta,0) = \begin{cases} 0, & \theta \in \Theta_0 \\ 1, & \theta \in \Theta_1 \end{cases}$$

$$L(\theta,1) = \begin{cases} 0, & \theta \in \Theta_1 \\ 1, & \theta \in \Theta_0 \end{cases}$$

于是决策函数 $\delta(X)$ 的函数为

$$\begin{aligned} R(\theta,\delta(X)) &= \mathrm{E}_{X|\theta}L(\theta,\delta(X)) \\ &= \int_{\delta(x)=1} L(\theta,\delta(x))\mathrm{d}P_\theta + \int_{\delta(x)=0} L(\theta,\delta(x))\mathrm{d}P_\theta \\ &= \int_W L(\theta,1)\mathrm{d}P_\theta + \int_{\overline{W}} L(\theta,0)\mathrm{d}P_\theta \\ &= \begin{cases} P_\theta(x \in W), & \theta \in \Theta_0 \\ P_\theta(x \notin W), & \theta \in \Theta_1 \end{cases} \end{aligned}$$

这表明：当 $\theta \in \Theta_0$ 时，其风险就是犯第 Ⅰ 类错误的概率. 当 $\theta \in \Theta_1$ 时，其风险就是犯第 Ⅱ 类错误的概率，假如记 $\delta(x)$ 的势函数为

$$\beta_\delta(\theta) = P_\theta(X \in W), \quad \theta \in \Theta_0 \cup \Theta_1$$

那么风险函数与势函数间有如下关系

$$R(\theta,\delta(x))=\begin{cases}\beta_\delta(\theta), & \theta\in\Theta_0\\ 1-\beta_\delta(\theta), & \theta\notin\Theta_0\end{cases}$$

Neyman-Pearson 假设检验理论的基本思想是在限制犯第Ⅰ类错误概率不超过某一个正数 α 的条件下，寻找使犯第Ⅱ类错误概率尽可能小的拒绝域，这在统计决策理论中等价于寻找这样决策函数 $\delta^*(x)$，在满足

$$\sup_{\theta\in\Theta_0}R(\theta,\delta^*(x))\leqslant\alpha$$

条件下，使得对样本空间 $\mathscr{X}$ 中任一子集的示性函数 $\delta(x)$ 有

$$R(\theta,\delta^*(x))\leqslant R(\theta,\delta),\quad\forall\,\theta\in\Theta_1$$

所以寻找一致最优势检验的问题仍是特定损失函数下的统计决策问题. 第三章叙述的 Neyman-Pearson 假设检验理论表明，对单边假设检验问题，一致最优势检验有时要把决策函数扩大到随机化决策函数类中才能找到，而对双边假设检验问题，一致最优势检验要把决策函数限于无偏检验类中才能找到.

至此已看到统计推断中的三种基本形式都可看作特定行动空间和特定损失函数下的统计决策问题. 所以在统计决策理论框架里不仅能讨论经典统计量中的一些问题. 还可进一步研究很多新的有趣问题. 此外，我们还看到，在样本空间 $\mathscr{X}$ 到行动空间 Δ 上的一切决策函数组成的决策函数类 $\mathscr{D}$ 中寻找一致最优决策函数往往不能奏效，只能在其一个特定的子类 $\mathscr{D}_1\subset\mathscr{D}$ 中才有可能寻得一致最优决策函数. 或者扩大到随机化决策函数类中才能寻得一致最优决策函数.

§5.2.4　最小最大估计

假如我们不是全面对风险函数进行逐点比较，而只对风险函数某一侧面进行比较，从中选出在这一侧面上最优的决策函数. 这就形成了众多的统计决策准则. 其中尤以最小最大准则和 Bayes 风险准则最常用和最有意义. 这里先叙述最小最大准则.

设有一可控参数结构$(\mathscr{X},\mathscr{B},\{p_\theta:\theta\in\Theta\})$，行动空间 Δ 和损失函

数 $L(\theta,\delta)$，又设 $\mathscr{D}$ 为该统计决策问题的某个决策函数类，其风险函数如(5.18)式所示.

定义 5.5 在上述假设下，若在决策函数类 $\mathscr{D}$ 中可选出这样的决策函数 $\delta^*(x)$，使得

$$\sup_{\theta\in\Theta} R(\theta,\delta^*(x)) = \inf_{\delta(x)\in\mathscr{D}} \sup_{\theta\in\Theta} R(\theta,\delta(x)) \tag{5.19}$$

则称 $\delta^*(x)$ 为该统计决策问题在最小最大风险准则下的最优决策函数，或称最大风险最小化的决策函数，简称最小最大(Minimax)决策函数，在点估计问题中，$\delta^*(x)$ 还称为最小最大(Minimax)估计. 相应的风险称为最小最大风险.

最小最大决策函数的获得可分为两步进行. 第一步是对 $\mathscr{D}$ 中每个决策函数 $\delta(x)$ 算出其最大风险值 $\sup_{\theta\in\Theta} R(\theta,\delta(x))$；第二步是在所有的最大风险值中选取相对最小值. 此值对应的决策函数便是 Minimax 决策函数. 从上述求解过程可见，最小最大准则是把风险函数的比较转化为最大风险比较，是在所有最大风险中选择具有最小风险的决策函数. 从策略观点上来看，它是预防最大风险出现的一种稳妥策略，一种保守的策略，这种策略，遭到不少人的批评. 譬如有两个决策函数 $\delta_1(x)$ 和 $\delta_2(x)$，其风险函数(见图 5.2(b))一个为常数，另一个在大多数场合都有较小的风险，只有很少机会出现较大风险. 有些人主张选用 $\delta_2(x)$. 可另一些人主张选用 $\delta_1(x)$. 这种差别大多是由于决策人所处的环境不同而引起的. 譬如，当企业较小，资金薄弱，经济上经不起大的冲击时，常采用最小最大准则. 又如很多人参加人寿保险和财产保险，也出于此种策略，他们宁可平时无事花点钱，免招突然的灭顶之灾.

例 5.9 在例 5.8 中我们遇到有 8 个决策函数组成类

$$\mathscr{D} = \{\delta_1,\delta_2,\delta_3,\delta_4,\delta_5,\delta_6,\delta_7,\delta_8\}$$

假如参数空间 Θ 选为[0,0.12]，即不合格品率 θ 在 0 到 0.12 之间，那从表 5.1 上容易看出 $\mathscr{D}$ 中每个决策函数所具有的最大风险值(见表 5.2)，从中立即可得 $\delta_5(x)$ 是最小最大风险准则下的最优决策函数. 其中 $\delta_5(x)$ 为

$$\delta_5(x) = \begin{cases} \delta_2, & x = 0 \\ \delta_1, & x = 1,2 \end{cases}$$

即从每箱中随机抽取两个产品进行检查，当其中没有一件是不合格品时就不再检查其它产品；当其中至少有一件是不合格品时就对全箱产品逐一检查. 采用 $\delta_5(x)$ 所引起的最大平均损失是 55.45 元/箱. 这是它的最小最大风险.

表 5.2 8 个风险函数的最大值

$\delta(x)$	$\sup R(\theta,\delta(x))$	$\delta(x)$	$\sup R(\theta,\delta(x))$
$\delta_1(x)$	78.4	δ_5	55.45
$\delta_2(x)$	78.4	δ_6	56.48
$\delta_3(x)$	78.4	δ_7	70.57
$\delta_4(x)$	78.4	δ_8	71.60

例 5.10 设 X 是正态总体 $N(\theta,1)$ 抽取的(容量为 1)样本.

$$\mathscr{D}_1 = \{\delta_c : \delta_c(X) = cX, c\ \text{为任意实数}\}$$

寻找 θ 的估计. 若取平方损失，则 $\mathscr{D}_1$ 中任一个估计 δ_c 的风险函数为

$$R(\theta,\delta_c) = \mathrm{E}_{x|\theta}(cX-\theta)^2 = \mathrm{E}_{x|\theta}[c(X-\theta)+(c-1)\theta]^2$$
$$= c^2 + (c-1)^2\theta^2$$

容易看出，当 $c=1$ 时，$R(\theta,\delta_1)=1$. 而当 $c>1$ 时，有

$$R(\theta,\delta_1) < R(\theta,\delta_c), \quad \forall\ \theta \in \mathbf{R}$$

所以当 $c>1$ 时，δ_1 是一致优于 δ_c. 可当 $c\leqslant 1$ 时，诸 δ_c 的风险函数呈交叉状态，故对诸风险函数逐点比较，在 $\mathscr{D}_1$ 中没有一致最小风险估计. 但它们在 $\mathbf{R}$ 上的最大风险可以算出

$$\sup_{\theta\in\mathbf{R}} R(\theta,\delta_c) = \sup_{\theta\in\mathbf{R}}[c^2+(c-1)^2\theta^2] = \begin{cases}1, & c=1\\ \infty, & c\neq 1\end{cases}$$

可见，按最小最大准则，$\delta_1(X)=X$ 是 θ 在 $\mathscr{D}_1$ 中的最小最大估计，其最小最大风险为 1.

最后指出，至今尚无直接获得最小最大决策函数(估计)的简单方法，但有一些间接方法可以采用，这将在以后有关章节中再加以叙述.

§5.2.5 随机化决策函数

当给定样本观察值 x 时,非随机化决策函数 $\delta(x)$ 在 Δ 上选取行动是唯一确定的.可在某些统计决策问题(如假设检验问题)中,即使样本观察值 x 给定,还有必要以某种随机方式在 Δ 中选取行动才能改善决策效果.假如这里的随机方式用概率分布明确表示,那么就获得随机化决策函数.

定义 5.6 在给定的统计决策问题中,若对样本空间 $\mathscr{X}$ 中每个 x 在行动空间 $(\Delta,\mathscr{B}_\Delta)$ 上有一个条件分布 $\delta(\cdot|x)$,则条件分布族

$$\{\delta(D|x),\forall D\in\mathscr{B}_\Delta: x\in\mathscr{X}\} \tag{5.20}$$

称为该统计决策问题中的一个随机化决策函数,记为 $\delta(D|x)$.该统计决策问题中的一切随机化决策函数组成的类称为随机化决策函数类,记为 $\overline{\mathscr{D}}$.

从上述定义可以看出,这个随机化决策函数类 $\overline{\mathscr{D}}$ 是一个很大的类,它包括一切可能定义在 $(\Delta,\mathscr{B}_\Delta)$ 上的概率分布.当然也包括一切可能的退化分布.当 $\delta(D|x)$ 是退化分布,则在给定 x 下,随机化决策函数将以概率 1 在 Δ 中取某个行动,记这个行动为 $\delta(x)$,这样一来,$\delta(D|x)$ 就是非随机化决策函数 $\delta(x)$.从而有:$\overline{\mathscr{D}}\supset\mathscr{D}$,其中 $\mathscr{D}$ 为同一统计决策问题中的非随机化决策函数类

从上述定义还可看出,假如把非随机化决策函数 $\delta(x)$ 看作为"一次抽样"的话,那么随机化决策函数 $\delta(D|x)$ 可看作为"二次抽样",这二次分别是:

第一次:在样本空间 $\mathscr{X}$ 中按总体分布 $p_\theta(x)$ 随机抽取一个样本 x_0,与此同时也就确定一个条件分布 $\delta(D|x_0)$;

第二次:在行动空间 Δ 中按条件分布 $\delta(D|x_0)$ 抽取一个行动,用此行动对自然界(或社会)的状态(或参数)作出判断.为了实现这一点,可以按条件分布 $\delta(D|x_0)$ 设计一个随机试验或随机模拟,使其产生的一个结果就等价于从 $\delta(D|x_0)$ 中随机抽取一个样本.

例 5.11 设有一个统计结构 $(\mathbf{R}^n,\mathscr{B}_{\mathbf{R}^n},\{p_\theta:\theta\in\Theta\})$ 和一个行动空间 $(\Delta,\mathscr{B}_\Delta)$,假如行动空间 Δ 只含有两个元素,那么相应的随机化决

策函数 $\delta(D|x)$ 就是 Δ 上的二点分布族. 对于这类分布只要对 Δ 中某一个元素给出概率, 那么此二点分布就完全确定了. 譬如, 记 $\Delta=\{a_0, a_1\}$, 对任意的 $x\in\mathscr{X}$, 给出 a_1 的概率

$$\delta(a_1|x)=\phi(x),\quad 0\leqslant\phi(x)\leqslant 1$$

那么

$$\delta(a_0|x)=1-\phi(x)$$

这样一来, 只要在样本空间 $\mathscr{X}$ 上给出一个仅在 $[0,1]$ 上取值的函数 $\phi(x)$, 就可以构造一个随机化决策函数. 假设检验理论中正是这样做的. 在那里称上述定义中的 $\phi(x)$ 为检验函数, 特别当 $\phi(x)$ 只能取 0 或 1 时, 即 $\phi(x)$ 是某个集的示性函数时, 这种 $\phi(x)$ 对应的决策函数就是非随机化的.

类似地, 当 $\Delta=\{a_1,a_2,\cdots\}$ 含有有限个或可数个行动时, 只要在 $\mathscr{X}$ 上给出有限个或可数个实函数 $\phi_1(x),\phi_2(x),\cdots$, 使得

(1) $\phi_i(x)\geqslant 0,\quad i=1,2,\cdots,$

(2) $\sum_{i=1}^{\infty}\phi_i(x)=1,$

就可以构造一个随机化决策函数.

$$\delta(a_i|x)=\phi_i(x),\quad i=1,2,\cdots,x\in\mathscr{X}$$

随机化决策函数主要在理论研究中使用, 出于某种优良性而需要引入随机化决策函数. 但它在实际中很少被采用. 因为把行动的最后决策留给某种随机化方法来决定对当今的决策者是难以接受的. 但在遇到聪明的对手时, 随机化决策是最好的方法. 美国统计学家 Berger 对此举了如下的例子[2].

例 5.12 甲乙二人进行赌博, 二人同时亮出手中的一枚硬币. 若两枚硬币同为正面或同为反面, 则甲得 1 元, 否则乙得 1 元. 由于甲对正反两面任何有规律的选取方法都会被聪明的对手(乙)识破, 从而获胜. 甲为了避免这种根本性的失败的唯一方法就是甲按随机方式选取正面或反面. 因为只有这样乙的聪明才智无法施展. 赌博在机会均等下进行.

§5.2.6 随机化决策函数的风险函数

如何评价一个随机化决策函数 $\delta(D|x)$ 的优劣呢？大家知道，一个用 $\delta(D|x)$ 产生的行动 a 要经过二次抽样才可获得，一是从总体分布 $p_\theta(x)$ 随机抽取一个样本 x，二是从条件分布 $\delta(D|x)$ 随机抽取一个行动 a，这样获得的行动具有二重随机性. 当把这个行动代入损失函数 $L(\theta,a)$ 中时，损失函数也有二重随机性. 为了消除这二重随机性的影响，我们要作二次平均才能确定一个非随机量. 下面的定义正是这样考虑的.

定义 5.7 设有一个参数结构 $(\mathscr{X},\mathscr{B},\{P_\theta:\theta\in\Theta\})$ 和一个行动空间 $(\Delta,\mathscr{B}_\Delta)$，又设 $\delta(D|x)$ 是其上的一个随机化决策函数，则把损失函数 $L(\theta,a)$ 对条件分布 $\delta(D|x)$ 的数学期望

$$\overline{L}(\theta,\delta)=\mathrm{E}_{a|x}[L(\theta,a)]=\int_\Delta L(\theta,a)\delta(\mathrm{d}a|x) \tag{5.21}$$

称为 $\delta(D|x)$ 的加权损失函数，而把 $\overline{L}(\theta,\delta)$ 对总体分布 $p_\theta(x)$ 的数学期望

$$\begin{aligned}R(\theta,\delta)&=\mathrm{E}_{x|\theta}\{\mathrm{E}_{a|x}[L(\theta,a)]\}\\&=\int_{\mathscr{X}}\overline{L}(\theta,\delta)p_\theta(x)\mathrm{d}x\\&=\int_{\mathscr{X}}\left\{\int_\Delta L(\theta,a)\delta(\mathrm{d}a|x)\right\}p_\theta(x)\mathrm{d}x\end{aligned} \tag{5.22}$$

称为 $\delta(D|x)$ 的风险函数

显然，$\delta(D|x)$ 的加权损失函数反应了在 $X=x$ 时，随机化决策函数 $\delta(D|x)$ 所引起的平均损失，这时 $\overline{L}$ 仅是 θ 和样本观察值 x 的函数，特别当 $\delta(D|x)$ 为退化分布时，有 $\overline{L}(\theta,\delta)=L(\theta,\delta)$，这意味着如上定义的加权损失函数既保持了原来损失函数的特性，又能反映在随机化决策函数场合引起的损失函数，从而加权损失函数对 $p_\theta(x)$ 的平均所得到的风险函数，也包含了原来风险的各种特性. 因此用定义 5.7 中的风险函数作为评价一个随机化决策函数的尺度是合理的，也与非随机化决策函数的评价是统一的.

在给定的统计决策问题中，把决策函数类从非随机化决策函数类 $\mathscr{D}$ 拓广到随机化决策函数类 $\overline{\mathscr{D}}$ 时，决策效果往往可以得到改善. 这一点在假设检验理论研究中已得到证实. 在那里由于引入随机化检验，致使一致最优势检验(UMPT)存在. 类似地在 $\overline{\mathscr{D}}$ 上寻求最小最大估计要比在 $\mathscr{D}$ 上的寻求最小最大估计在最小最大风险上会得到改善. 下面的例子说明这个问题.

例 5.13[8]　设随机变量 X 服从 Poisson 分布 $P(\lambda)$，其中未知参数 λ 仅可能取 $\lambda_1=1$ 和 $\lambda_2=2$ 中的一个. 根据一次试验结果 x，构造如下十个统计量.

$$\delta_1(x)=\lambda_2, x=0,1,\cdots$$

$$\delta_2(0)=\lambda_1;\delta_2(x)=\lambda_2, x=1,2,\cdots$$

$$\delta_3(x)=\lambda_1, x=0,1;\delta_3(x)=\lambda_2, x=2,3,\cdots$$

$$\delta_4(x)=\lambda_1, x=0,1,2;\delta_4(x)=\lambda_2, x=3,4,\cdots$$

$$\delta_5(x)=\lambda_1, x=0,1,2,3;\delta_5(x)=\lambda_2, x=4,5,\cdots$$

$$\delta_6(x)=\lambda_1, x=0,1,2,3,4;\delta_6(x)=\lambda_2, x=5,6,\cdots$$

$$\delta_7(x)=\lambda_1, x=0,1,\cdots$$

$$\delta_8(x)=\lambda_1, x\text{ 为偶数};\delta_8(x)=\lambda_2, x\text{ 为奇数}$$

$$\delta_9(x)=\lambda_1, x=1,2,\cdots;\delta_9(0)=\lambda_2$$

$$\delta_{10}(x)=\lambda_1, x=2,3,\cdots;\delta_{10}(x)=\lambda_2, x=0,1$$

在这 10 个估计中，前几种是合理的，而后几种纯粹是数学上为说明问题而设置的，从实用观点是不合理的. 这不影响我们按合理准则挑选最优的估计.

在这个问题中，行动空间 $\Delta=\{a\}$ 取得与参数空间 $\Theta=\{\lambda\}$ 相同，即 $\Delta=\Theta=\{1,2\}$. 而取损失矩阵如下：

$$\begin{array}{cc} a=1 & a=2 \end{array}$$

$$L=\begin{bmatrix} 0 & 50 \\ 100 & 0 \end{bmatrix}\begin{array}{l} \lambda=1 \\ \lambda=2 \end{array}$$

在这样一个统计决策问题中，上述 10 个估计的各种风险函数值

$R(\lambda_i, a_j)$, $i=1,2$; $j=1,2,\cdots,10$. 计算结果列于表 5.3 中.

表 5.3 10 个风险函数在 $\lambda=1,2$ 的值

δ_j	$R(\lambda_1,\delta_j)$	$R(\lambda_2,\delta_j)$	$\max_i R(\lambda_i,\delta_j)$
δ_1	50	0	50
δ_2	31.6	13.5	31.6(min)
δ_3	13.2	40.6	40.6
δ_4	4.02	67.7	67.7
δ_5	0.95	85.7	85.7
δ_6	0.18	94.7	94.7
δ_7	0	100.0	100.0
δ_8	21.6	50.9	50.9
δ_9	14.8	86.5	86.5
δ_{10}	36.8	67.7	67.7

从表 5.3 上可见，当 $\mathscr{D}_1$ 是上述 10 个估计组成时，参数 λ 的最小最大估计是 $\delta_2(x)$. 假如把随机化估计也作为考虑对象，那么参数 λ 的最小最大估计将会得到改善. 下面分几步来研究这个问题.

1. 每一个估计量 δ_j 在 λ_1 和 λ_2 各有一个风险值，记为 $y_1=R(\lambda_1,\delta_j)$ 和 $y_2=R(\lambda_2,\delta_j)$. 假如把 y_1,y_2 看作是直角坐标系的两个坐标，那么 $\mathscr{D}_1$ 中每一个估计量就对应此坐标系中的一点. 图 5.3 作出了 $\delta_1,\cdots,\delta_{10}$所对应的点. 从这种几何解释可以看到，估计类 $\mathscr{D}_1$ 是由第一象限内的 10 个点组成的一个集合. 假如在 $\mathscr{D}_1$ 中增加一个估计，就相当于在坐标平面上增加一点.

2. 我们把随机化估计增加到 $\mathscr{D}_1$ 中去，为此先考虑一种特殊形式的随机化估计量 δ^*，它在样本 x 给定条件下以概率 p 取 $\delta_2(x)$，而以概率 $1-p$ 取 $\delta_3(x)$，其中 $0<p<1$，也可记为：

$$\begin{array}{c|cc} \delta^* & \delta_2(x) & \delta_3(x) \\ \hline P & p & 1-p \end{array} \tag{5.23}$$

考虑到 $\delta_2(x)$ 和 $\delta_3(x)$ 的具体定义，δ^* 有如下含义：

$$\delta^*(0) = \lambda_1$$

$\delta^*(1)$ 是一个随机变量，其分布为：

$$P(\delta^*(1) = \lambda_2) = p$$

$$P(\delta^*(1) = \lambda_1) = 1 - p$$

$$\delta^*(x) = \lambda_2, x = 2, 3, \cdots$$

这种形式的随机化估计的风险函数可以从式子(5.22)求得，但在这里它可以通过 $\delta_2(x)$ 和 $\delta_3(x)$ 的风险函数的加权平均来计算. 这是因为，对 $i=1,2$，有

$$\begin{aligned}
& R(\lambda_i,\delta_2)p + R(\lambda_i,\delta_3)(1-p) \\
&= \sum_x L(\lambda_i,\delta_2)P_{\lambda_i}(x)p + \sum_x L(\lambda_i,\delta_3)P_{\lambda_i}(x)(1-p) \\
&= \sum_x [L(\lambda_i,\delta_2)p + L(\lambda_i,\delta_3)(1-p)]P_{\lambda_i}(x) \\
&= \sum_x L(\lambda_i,\delta^*(x))P_{\lambda_i}(x) \\
&= R(\lambda_i,\delta^*)
\end{aligned}$$

由此可以看出，这种随机化估计量 $\delta^*(x)$ 在坐标平面上对应的点是位于在 δ_2 和 δ_3 的连线上，p 的不同只意味着它在连线上的位置不同，但都一定在连线之间.

这个例子是具有一般性的，可以想象，假如把形如(5.23)的随机化估计都加入到 $\mathscr{D}_1$ 内，则在坐标平面上就等于增加原来 10 个点之间的两两连线. 假如在这些线段上任取两点，作形如(5.23)的随机化估计量，并把这些估计量也加入到 $\mathscr{D}_1$ 中去，那么在坐标平面上的点就形成一个区域，记为 $\overline{\mathscr{D}}_1$，并且还是一个凸区域. 这是因为形成这个区域的过程就是按照凸区域的要求进行的.

3. 最后所得凸区域 $\overline{\mathscr{D}}_1$ 上的点可以分为两类. 一类是此凸区域的“西南方向”的边界点，即靠近 y_1 轴和 y_2 轴的边界点；其他的点组成第二类点. 显然，对第二类中任一点 (y'_1, y'_2)，可以在第一类中找到一点 (y_1, y_2)，使得 $y_1 < y'_1$，$y_2 < y'_2$. 这只要连接原点与 (y'_1, y'_2)，其连线与“西南方向”边界线的交点就是所要找的点. 因此在凸区域上寻求参数

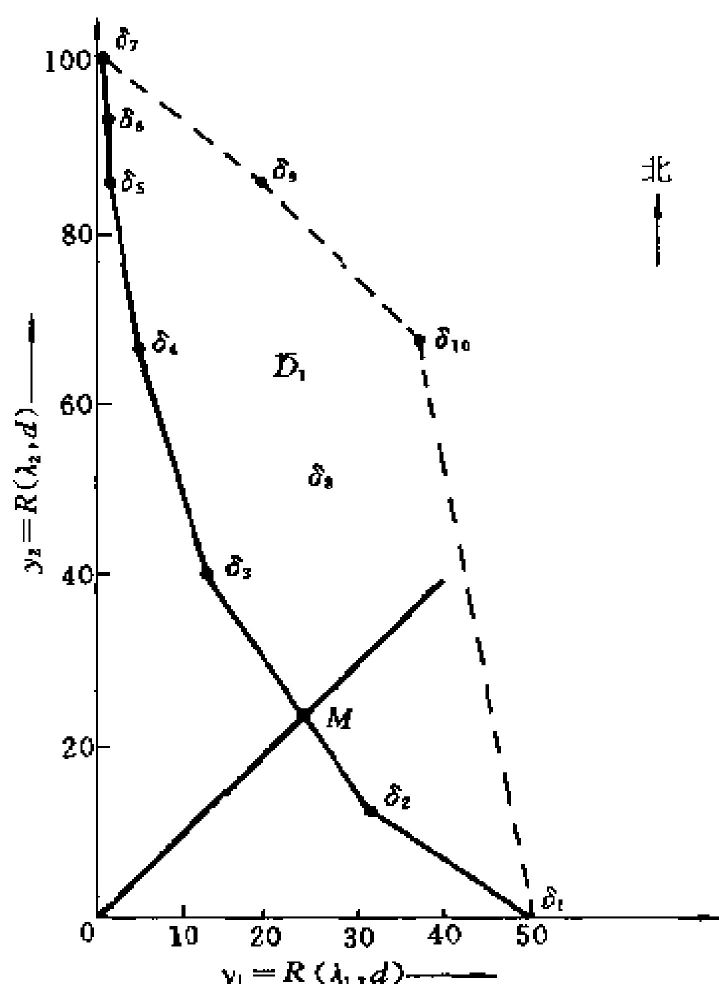


图 5.3 λ 的估计量的几何表示(例 5.13)

λ 的最小最大估计,可以仅限于在第一类点中寻找.

4. 为了在 $\mathscr{D}_1$ 中寻找 λ 的最小最大估计,我们借助于直观的几何作图方法,在图 5.3 上作直线 $y_2=y_1$,它与西南边界交于 M 点,则点 M 对应的估计就是所要寻找的最小最大估计. 这是因为在直线 OM 下方的点有 $y_1>y_2$;在直线 OM 上方的点有 $y_1<y_2$;在直线 OM 上的点有 $y_1=y_2$. 所以"西南方向"边界线上任一点(除点 M 外),至少有一个坐标要大于点 M 的坐标. 注意到这些坐标都是风险函数的取值,就可说明点 M 对应的估计就是所要寻找的参数 λ 的最小最大估计.

由于点 M 在线段 $\delta_2\delta_3$ 上,故对应的随机化估计量 $\delta^*(x)$具有形式(5.23),其中关键在于寻求 p. 容易求得点 M 的两个坐标皆为 24.28,于是利用风险函数间的关系式(5.23)可得方程:

$$31.6p + 13.2(1-p) = 24.28$$

解之可得 $p=0.6$，这就说明参数 λ 的最小最大估计 $\delta^*(x)$ 是具有如下分布的随机化估计：

δ^*	$\delta_2(x)$	$\delta_3(x)$
P	0.6	0.4

注意到估计 $\delta_2(x)$ 和 $\delta_3(x)$ 的具体形式，还可以把 δ^* 写作：$\delta^*(0)=\lambda_1$；$\delta^*(1)$ 是一个随机变量，其分布为

$$P(\delta^*(1)=\lambda_2)=0.6, P(\delta_*(1)=\lambda_1)=0.4$$

$$\delta^*(x)=\lambda_2, x=2,3,\cdots$$

这个估计仅在 $x=1$ 时是随机的，而在 $x\neq 1$ 时都是非随机的. 根据 $x=1$ 时的分布设计一个随机试验是不难的. 这个随机化估计的最小最大风险为 24.28，要比原 $\mathscr{D}_1$ 中的最小最大估计 $\delta_2(x)$ 的最小最大风险 31.6 要小一些.

顺便指出，从这个例子可以看出，随机化决策函数类 $\overline{\mathscr{D}}$ 对凸组合运算是封闭的. 而非随机化决策函数类 $\mathscr{D}$ 对凸组合运算是不封闭的. 对凸组合运算都不封闭的决策函数类不能认为是满意的. 所以从数学观点看，决策函数研究不应限于非随机化决策函数类.

§5.3 决策函数的容许性

§5.3.1 决策函数的容许性

利用风险函数对决策函数的容许性作出评价不在于选优，而着眼于除劣.

定义 5.8 对给定的统计决策问题和随机化决策函数类 $\overline{\mathscr{D}}$，决策函数 $\delta_0(D|x)$ 称为非容许的，假如在 $\overline{\mathscr{D}}$ 中存在有另一个决策函数 $\delta_1(D|x)$ 满足如下两个条件：

1. $R(\theta,\delta_1)\leqslant R(\theta,\delta_0),\forall\ \theta\in\Theta$

2. 在 Θ 中至少有一个 θ_0，有 $R(\theta_0,\delta_1)<R(\theta_0,\delta_0)$

假如在 $\mathscr{D}$ 中不存在满足上述二条件的决策函数，则称 $\delta_0(D|x)$ 为容许的.

从这个定义可见，非容许决策函数是不应被采用的. 譬如，在凸损失函数下，随机化决策函数常常是非容许的(见定理 5.1). 若在 $\mathscr{D}$ 中剔去非容许决策函数后，余下的容许决策函数仍然是很多的，它们的风险函数呈交叉状态，各自在不同的地方比其他的要好一些. 所以容许决策函数还不一定是一个很有效的估计. 我们使用的决策函数至少应是容许的，但不是任一容许决策函数都可使用. 譬如，不依赖于样本的常数值估计 $\delta(x)=\theta_0$ 常常是容许估计，因为它在 $\theta=\theta_0$ 处给出精确的估计值. 这种只顾 θ_0 一点，不顾其他点的估计是不会为人们采用的. 决策函数的容许性的实际意义在于缩小挑选的范围.

容许性问题常在点估计问题中讨论，这一章也将主要讨论点估计的容许性问题.

例 5.14 设 $X_1,\cdots,X_n$ 是来自正态总体 $N(0,\sigma^2)$ 的一个样本. 大家知道，常用的估计(这里是在非随机化决策函数里讨论)

$$\hat{\sigma}_1^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2,\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

是 σ^2 的无偏估计，且在平方损失函数下它的风险 $R(\sigma^2,\hat{\sigma}_1^2)$ 为

$$\mathrm{E}(\hat{\sigma}_1^2-\sigma^2)^2=\frac{2\sigma^4}{n-1}$$

假如我们考虑形如 $\delta(x)=c\hat{\sigma}_1^2$ 的估计，在同样的二次损失函数下，$\delta(x)$ 的风险为

$$\begin{aligned}R(\sigma^2,\delta(x))&=\mathrm{E}(c\,\hat{\sigma}_1^2-\sigma^2)^2\\&=c^2\mathrm{E}(\hat{\sigma}_1^2-\sigma^2)^2+\sigma^4(1-c)^2\\&=\sigma^4\left[\frac{2c^2}{n-1}+(1-c)^2\right]\end{aligned}$$

令 $g(c)=\dfrac{2c^2}{n-1}+(1-c)^2$，它是 c 的二次函数，且在 $c_0=\dfrac{n-1}{n+1}$ 处达到最小，所以令

$$\hat{\sigma}_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

则有

$$R(\sigma^2, \hat{\sigma}_1^2) > R(\sigma^2, \hat{\sigma}_2^2) = \frac{2}{n+1}\sigma^4$$

这样就证明了估计 $\hat{\sigma}_1^2$ 是非容许的，因为 $\hat{\sigma}_2^2$ 比 $\hat{\sigma}_1^2$ 的风险一致更小. 但要注意，这并不意味着估计 $\hat{\sigma}_2^2$ 一定是容许的，譬如若取

$$\hat{\sigma}_3^2 = \frac{1}{n+2}\sum_{i=1}^{n}x_i^2$$

它的风险

$$R(\sigma^2, \hat{\sigma}_3^2) = \mathrm{E}(\hat{\sigma}_3^2 - \sigma^2)^2 = \frac{2}{n+2}\sigma^4$$

由于 $R(\sigma^2, \hat{\sigma}_3^2) < R(\sigma^2, \hat{\sigma}_2^2)$，所以 $\hat{\sigma}_2^2$ 仍是非容许估计.

§5.3.2 Stein 效应

正态均值用其样本均值去估计有很好的性质，它是无偏的，方差最小的和有效的. 当把这样的估计推广到 p 元正态分布场合时出现了意想不到的结果. Stein 在 1955 年指出，在二次损失函数下，$p \geq 3$ 时，样本均值向量是正态均值向量的非容许估计. 这个结果在当时统计界引起轰动效应，产生了研究容许性的热潮，并持续二、三十年之久. 如今把这个效应称为 Stein 效应(Stein effect). 这里先用例子形式来叙述 Stein 的结果，然后叙述其影响.

例 5.15 p 元正态总体均值的估计.

设 $\boldsymbol{X} = (X_1, \cdots, X_p)' \sim N_p(\boldsymbol{\mu}, \boldsymbol{I}_p)$，其中 $\mu = (\mu_1, \cdots, \mu_p)' \in \mathbf{R}^p$. 如今对 $\boldsymbol{X}$ 作一次观察，并用观察结果 $\boldsymbol{\delta}(\boldsymbol{X}) = (X_1, \cdots, X_p)$ 去估计 μ. 现在二次损失函数

$$L(\mu, \boldsymbol{\delta}) = (\boldsymbol{\delta} - \boldsymbol{\mu})'(\boldsymbol{\delta} - \boldsymbol{\mu})$$

下研究 $\boldsymbol{\delta}(\boldsymbol{X})$ 的容许性问题. Stein 在 1955 年指出 $\boldsymbol{\delta}(\boldsymbol{X})$ 在 $p \geq 3$ 时是 $\boldsymbol{\mu}$ 的非容许估计[9]，但证明不是结构性的，到 1961 年，James 和 Stein 给出了比 $\boldsymbol{\delta}(\boldsymbol{X})$ 一致更优的估计[10].

$$\delta^{JS}(\boldsymbol{X})=\left(1-\frac{p-2}{\boldsymbol{X}'\boldsymbol{X}}\right)X \tag{5.24}$$

这个估计被称为 James-Stein 估计. 选用这个估计的直观想法出自于

$$\mathrm{E}(\boldsymbol{X}'\boldsymbol{X})=\mathrm{E}(X_1^2+\cdots+X_p^2)=p+\boldsymbol{\mu}'\boldsymbol{\mu}$$

这就告诉我们,当用 $\boldsymbol{X}$ 去估计 $\boldsymbol{\mu}$ 时,$\boldsymbol{X}$ 的平均长度 $\mathrm{E}(\boldsymbol{X}'\boldsymbol{X})$实际上比 $\boldsymbol{\mu}$ 的长度大. 这是一种系统偏差,需要改进. 改进的方法就是将 $\boldsymbol{X}$ 乘以某一个修正因子. Stein 考虑到这个修正因子与 $\boldsymbol{X}$ 和 p 有关,故选用 $\left(1-\dfrac{p-2}{\boldsymbol{X}'\boldsymbol{X}}\right)$作为修正因子.

为了证明 $\delta(\boldsymbol{X})$是 $\boldsymbol{\mu}$ 的非容许估计,我们只要证明

$$R(\boldsymbol{\mu},\boldsymbol{\delta}(x))-R(\boldsymbol{\mu},\boldsymbol{\delta}^{JS}(x))>0$$

即可. 为此需要进行一系列计算

$$\begin{aligned}
&R(\boldsymbol{\mu},\boldsymbol{\delta}(x))-R(\boldsymbol{\mu},\boldsymbol{\delta}^{JS}(x))\\
&=\mathrm{E}(\boldsymbol{X}-\boldsymbol{\mu})'(\boldsymbol{X}-\boldsymbol{\mu})-\mathrm{E}\left[\boldsymbol{X}-\frac{p-2}{\boldsymbol{X}'\boldsymbol{X}}\boldsymbol{X}-\boldsymbol{\mu}\right]'\left[\boldsymbol{X}-\frac{p-2}{\boldsymbol{X}'\boldsymbol{X}}\boldsymbol{X}-\boldsymbol{\mu}\right]\\
&=(p-2)\left\{2-2\mathrm{E}\left(\frac{\boldsymbol{\mu}'\boldsymbol{X}}{\boldsymbol{X}'\boldsymbol{X}}\right)-(p-2)\mathrm{E}\left(\frac{1}{\boldsymbol{X}'\boldsymbol{X}}\right)\right\}
\end{aligned}$$

下面我们将来计算上式花括号中的两个数学期望.

我们知道,当 $\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\boldsymbol{I}_p)$时,$\boldsymbol{X}'\boldsymbol{X}$ 服从非中心 Gamma 分布 $Ga(\alpha,\lambda,\gamma)$,其中 $\alpha=\dfrac{p}{2}$,$\lambda=\dfrac{1}{2}$,$\gamma=\dfrac{1}{2}\sum_{i=1}^{p}\mu_i^2$,所以

$$\begin{aligned}
\mathrm{E}\left(\frac{1}{\boldsymbol{X}'\boldsymbol{X}}\right)&=\sum_{k=0}^{\infty}\frac{\mathrm{e}^{-\gamma}\gamma^k}{k!}\frac{1}{2^{\frac{p}{2}+k}\Gamma\left(\frac{p}{2}+k\right)}\int_0^{\infty}u^{\frac{p}{2}+k-2}\mathrm{e}^{-\frac{u}{2}}\mathrm{d}\mu\\
&=\sum_{k=0}^{\infty}\frac{\mathrm{e}^{-\gamma}\gamma^k}{k!}\frac{1}{p+2k-2}=\mathrm{E}_{\gamma}\left(\frac{1}{p+2K-2}\right)
\end{aligned}$$

假如把 K 看作一个随机变量,它服从参数为 $\gamma=\dfrac{1}{2}\sum_{i=1}^{p}\mu_i^2$ 的 Poisson 分布. 这样一来,上式数学期望就是对此 Poisson 分布求的. 下面我们来计算第二个期望,为此作如下正交变换:

$$\boldsymbol{Y}=\begin{pmatrix}Y_1\\ \vdots\\ Y_p\end{pmatrix}=\begin{pmatrix}\boldsymbol{\mu}'/\sqrt{\boldsymbol{\mu}'\boldsymbol{\mu}}\\ *\\ \vdots\\ *\end{pmatrix}\boldsymbol{X}$$

其中 * 表示与 $\boldsymbol{\mu}'/\sqrt{\boldsymbol{\mu}'\boldsymbol{\mu}}$ 正交的任一个行向量. 于是

$$\mathrm{E}(\boldsymbol{Y})=\left(\sqrt{\boldsymbol{\mu}'\boldsymbol{\mu}},0,\cdots,0\right)'\triangleq \boldsymbol{b}$$

$$\mathrm{Var}(\boldsymbol{Y})=\boldsymbol{I}_p$$

这表明 $\boldsymbol{Y}\sim N_p(\boldsymbol{b},\boldsymbol{I}_p)$，并且在上述正交变换下，有 $\boldsymbol{Y}'\boldsymbol{Y}=\boldsymbol{X}'\boldsymbol{X}$，$\boldsymbol{\mu}'\boldsymbol{X}=\sqrt{\boldsymbol{\mu}'\boldsymbol{\mu}}Y_1$，$Y_1$ 与 $(Y_2,\cdots,Y_p)$ 相互独立，由此可以看出 $(a=\sqrt{\boldsymbol{\mu}'\boldsymbol{\mu}})$

$$Y_1\sim N(a,1),\quad Z=\sum_{i=2}^{p}Y_i^2\sim Ga\left(\frac{p-1}{2},\frac{1}{2}\right)$$

所以 Y_1 与 Z 相互独立，其联合分布为

$$(Y_1,Z)\sim p(y_1)\cdot p(z)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{(y_1-a)^2}{2}}\frac{1}{2^{\frac{p-1}{2}}\Gamma\left(\frac{p-1}{2}\right)}z^{\frac{p-1}{2}-1}\mathrm{e}^{-\frac{z}{2}}$$

其中 $y_1\in\mathbf{R}$，$z\in\mathbf{R}^+$，于是

$$\mathrm{E}\left(\frac{\boldsymbol{\mu}'\boldsymbol{X}}{\boldsymbol{X}'\boldsymbol{X}}\right)=\mathrm{E}\left(\frac{aY_1}{\boldsymbol{Y}'\boldsymbol{Y}}\right)=a\mathrm{E}\left(\frac{Y_1}{Y_1^2+Z}\right)$$

$$=\frac{a}{\sqrt{2\pi}}\frac{\mathrm{e}^{-\frac{a^2}{2}}}{2^{\frac{p-1}{2}}\Gamma\left(\frac{p-1}{2}\right)}\int_{-\infty}^{\infty}\int_0^{\infty}\frac{y_1}{y_1^2+z}\mathrm{e}^{ay_1-\frac{y_1^2+z}{2}}z^{\frac{p-1}{2}-1}\mathrm{d}y_1\mathrm{d}z$$

作变换

$$u_1=y_1$$

$$u_2=y_1^2+z$$

其 Jacobian $|\boldsymbol{J}|=1$，于是

$$\mathrm{E}\left(\frac{\boldsymbol{\mu}'\boldsymbol{X}}{\boldsymbol{X}'\boldsymbol{X}}\right)=\frac{a\mathrm{e}^{-\frac{a^2}{2}}}{2^{\frac{p}{2}}\sqrt{\pi}\,\Gamma\left(\frac{p-1}{2}\right)}\cdot$$

$$\int_0^{\infty}\frac{\mathrm{d}u_2}{u_2}\int_{u_1^2<u_2}u_1(u_2-u_1^2)^{\frac{p-1}{2}-1}\mathrm{e}^{au_1-\frac{u_2}{2}}\mathrm{d}u_1$$

上述对 u_1 的积分可以利用下述积分公式：

$$\frac{1}{2^{\frac{p}{2}}\sqrt{\pi}\,\Gamma\left(\frac{p-1}{2}\right)}\int_{u_1^2<u_2} u_1(u_2-u_1^2)^{\frac{p-1}{2}-1}\mathrm{e}^{au_1-\frac{u_2^2}{2}}\mathrm{d}u_1$$

$$=a\sum_{k=1}^{\infty}\frac{\left(\frac{a^2}{2}\right)^{k-1}}{(k-1)!}u_2^{\frac{p}{2}+k-1}\mathrm{e}^{-\frac{u_2^2}{2}}\frac{1}{2^{\frac{p}{2}+k}\Gamma\left(\frac{p}{2}+k\right)}$$

代回原式，即得

$$\mathrm{E}\left(\frac{\mu'X}{X'X}\right)=a^2\mathrm{e}^{-\frac{a^2}{2}}\sum_{k=1}^{\infty}\frac{\left(\frac{a^2}{2}\right)^{k-1}}{(k-1)!}\frac{1}{2^{\frac{p}{2}+k}\Gamma\left(\frac{p}{2}+k\right)}\cdot$$

$$\int_0^{\infty}u_2^{\frac{p}{2}+k-2}\mathrm{e}^{-\frac{u_2^2}{2}}\mathrm{d}u_2$$

$$=a^2\mathrm{e}^{-\frac{a^2}{2}}\sum_{k=1}^{\infty}\frac{\left(\frac{a^2}{2}\right)^{k-1}}{(k-1)!}\frac{2^{\frac{p}{2}+k-1}\Gamma\left(\frac{p}{2}+k-1\right)}{2^{\frac{p}{2}+k}\Gamma\left(\frac{p}{2}+k\right)}$$

$$=\sum_{k=1}^{\infty}\frac{\left(\frac{a^2}{2}\right)^{k-1}}{k!}\mathrm{e}^{-\frac{a^2}{2}}\frac{2k}{p+2k-2}=\mathrm{E}_\gamma\left(\frac{2K}{p+2K-2}\right)$$

我们把 K 看作一个随机变量，它服从参数为 $\gamma=\frac{a^2}{2}=\frac{1}{2}\mu'\mu=\frac{1}{2}\sum_{i=1}^{p}u_i^2$ 的 Poisson 分布，这样一来，上式数学期望就是对此 Poisson 分布求得.

把上面算得两个数学期望代回原式，可得

$$R(\mu,\delta(x))-R(\mu,\delta^{JS}(x))$$

$$=(p-2)\left[2-2\mathrm{E}\left(\frac{\mu'X}{X'X}\right)-(p-2)\mathrm{E}\left(\frac{1}{X'X}\right)\right]$$

$$=(p-2)\left[2-2\mathrm{E}_\gamma\left(\frac{2K}{p+2K-2}\right)-(p-2)\mathrm{E}_\gamma\left(\frac{1}{p+2K-2}\right)\right]$$

$$=(p-2)^2\mathrm{E}_\gamma\left(\frac{1}{p+2K-2}\right)>0,\quad \text{当 } p>2$$

由此看出，当 $p \geqslant 3$ 时，有 $\delta(X)=X$ 是非容许估计.

这个结果发表以后，引起各国统计学家的思索，考虑了很多的问题. 如类似的现象是否在其它分布族中亦会出现吗？常用的统计量是否都是容许的？对一些分布族的参数，其容许估计的充分条件或充要条件是什么？这种现象是否是由于损失函数引起的？Stein 的这个结果给人们留下了极其深刻的印象，引起了一系列的研究.

譬如在研究 James-Stein 估计中，人们发现，当观察向量 x 接近于零时，因子 $[1-(p-2)/x'x]$ 会出现负值，甚至当 $x'x \to 0$ 时，这个因子会趋向 $-\infty$，对此 Baranchick (1970) 提出如下的截尾估计进行改进[11].

$$\delta_c^+(x)=\left(1-\frac{c}{x'x}\right)^+ x=\begin{cases}\left(1-\dfrac{c}{x'x}\right)x, & x'x>c\\ 0, & \text{其它}\end{cases} \tag{5.25}$$

其 c 满足 $p-2 \leqslant c \leqslant 2(p-2)$，这个估计被称为截尾 J-S 估计. 它一致地比不截尾 J-S 估计 (5.24) 要优，特别 δ_{p-2}^+ 对 δ^{JS} 作了改进. 但是对不同的 c 值，(5.25) 的风险函数仍不可比较，再想对截尾 J-S 估计作改进已很困难了. 对 $p=9$ 和 $c=2p-1$ 时 δ_c^+ 的风险函数如图 5.4 所示. 图中的水平线是 $\delta(X)=X$ 的风险函数 $R(\mu,\delta)=9$.

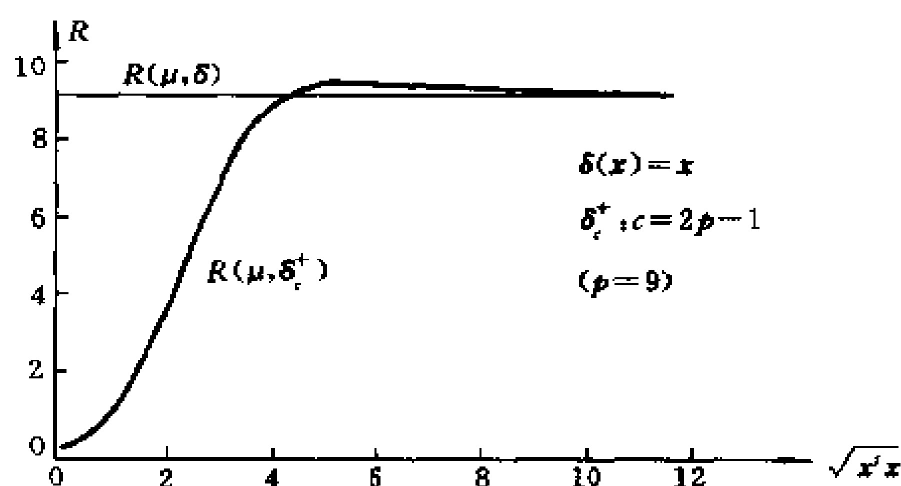


图 5.4　截尾 J-S 估计与均值的风险函数的比较

在 (5.24) 和 (5.25) 的基础上后来发展成一种"压缩估计"得到不少

应用. Cellier 等人(1989)还把 Stein 结果推广到球对称分布场合[12]. 在 Christian(1994)的书[11]中在这方面有大量叙述,并附有容许性方面较全面的文献.

§ 5.3.3 单参数指数族中的容许性问题

在 § 1.6 中给出单参数指数族的密度函数的标准形式

$$p_\theta(x) = \beta(\theta)\mathrm{e}^{\theta T(x)}h(x)$$

为了简化,我们把其中的 $h(x)$ 并入 Lebesgue 测度,并记 $\mathrm{d}\mu=h(x)\mathrm{d}x$ 这样一来,单参数指数族对 $\mathrm{d}\mu$ 的密度函数为

$$\frac{\mathrm{d}P_\theta}{\mathrm{d}\mu} = \beta(\theta)\mathrm{e}^{\theta T(x)}, \quad x \in \mathscr{X}, a < \theta < b \tag{5.26}$$

其中 $\beta(\theta)>0$, a 可取 $-\infty$, b 可取 ∞. 根据定理 1.14,充分统计量 $T(x)$ 的数学期望

$$\mathrm{E}[T(X)] = -\beta'(\theta)/\beta(\theta) \triangleq g(\theta) \tag{5.27}$$

Karlin(1958)在平方损失函数下给出参数 $g(\theta)$ 的容许估计的一种形式. 现不加证明地引述如下,其它分布族中的容许估计也有不少结果. 成平等人(1985)对此作了很好的总结[14],有兴趣的读者可参阅.

定理 5.1 设随机变量 X 服从指数分布(5.28),且

$$\int_{-\infty}^{\infty} \mathrm{e}^{\theta T(x)}\mathrm{d}\mu(x) < \infty, \qquad \theta \in (a,b) \tag{5.28}$$

若命 $a<c_1<c_2<b$,假如存在这样的 $\lambda\geqslant 0$,使得

$$\lim_{c_2\to b}\int_{c_1}^{c_2}\beta^{-\lambda}(\theta)\mathrm{d}\theta = \infty \tag{5.29}$$

$$\lim_{c_1\to a}\int_{c_1}^{c_2}\beta^{-\lambda}(\theta)\mathrm{d}\theta = \infty \tag{5.30}$$

那么在平方损失函数下,$\delta_\lambda(x)=T(x)/(1+\lambda)$是 $g(\theta)$的容许估计.

例 5.16 设随机变量 Y 服从 Gamma 分布 $Ga(\alpha, 1/\theta)$,它对 Lebesgue 测度 $\mathrm{d}y$ 的密度函数为

$$p(y;\theta)\mathrm{d}y = \frac{1}{\theta^\alpha\Gamma(\alpha)}y^{\alpha-1}\mathrm{e}^{-y/\theta}\mathrm{d}y, \quad 0 \leqslant y < \theta, 0 < \theta < \infty$$

其中参数 $\alpha>0$ 已知,θ 未知,要在平方损失下寻求 θ 的容许估计.

因为 $E(Y)=\alpha\theta$，我们取 $T(y)=y/\alpha, \tau=-\alpha/\theta$，于是 Gamma 分布相对于 $d\mu(y)=y^{\alpha-1}dy/(\alpha^{\alpha}\Gamma(\alpha))$ 的密度函数为

$$p(y;\tau)dy=\beta(\tau)e^{\frac{\tau}{\alpha}y}d\mu(y), \quad 0\leqslant y<\infty, -\infty<\tau<0$$

其中 $\beta(\tau)=(-\tau)^{\alpha}$. 容易验证，$E_{\tau}T(y)=-\beta'(\tau)/\beta(\tau)=-\dfrac{\alpha}{\tau}=\theta$，并且

$$\int_0^{\infty}e^{\frac{\tau}{\alpha}y}d\mu(y)<\infty$$

而对任意 $0<a<c<\infty$，有

$$\int_{-c}^{-a}\beta^{-\lambda}(\tau)d\tau=\int_{-c}^{-a}(-\tau)^{-\lambda\alpha}d\tau=\int_a^c u^{-\lambda\alpha}du$$

在 $c\to\infty$ 或 $a\to\infty$ 时，上述积分分别满足条件(5.29)和(5.30)，只要取 $\lambda=1/\alpha$. 根据定理 5.1，在平方损失函数下

$$\frac{T(y)}{\lambda+1}=\frac{y/\alpha}{1+1/\alpha}=\frac{y}{1+\alpha}$$

是 θ 的容许估计.

假如平方损失函数取 $L(\theta,d)=(d-\theta)^2/\theta^2$，那么 $y/(1+\alpha)$ 仍是 θ 的容许估计.

顺便指出，从上述讨论可知，在讨论估计的容许性，平方损失函数 $L(\theta,d)=\lambda(\theta)(d-\theta)^2$ 中 $\lambda(\theta)$ 取什么形式与估计的容许性无关.

利用上述结果还可以说明统计量 $\dfrac{1}{n+2}\sum_{i=1}^n x_i^2$ 是 σ^2 的容许估计，其中 $x_1,\cdots,x_n$ 是来自正态母体 $N(0,\sigma^2)$ 的一个样本.

大家知道，统计量 $Y=\sum_{i=1}^n x_i^2$ 是服从 $Ga\left(\dfrac{n}{2},\dfrac{1}{2\sigma^2}\right)$，它对Lebesgue 测度的密度函数为

$$p(y;\sigma^2)dy=\frac{1}{(2\sigma^2)^{\frac{n}{2}}\Gamma(n/2)}y^{\frac{n}{2}-1}e^{-y/2\sigma^2}dy,$$
$$0\leqslant y<\infty, 0<\sigma^2<\infty$$

它是 $\theta=2\sigma^2, \alpha=\dfrac{n}{2}$ 的 Gamma 分布，于是 $y/(1+\alpha)=2\sum_{i=1}^n x_i^2/(n+2)$ 是 $\theta=2\sigma^2$ 的容许估计. 为了求得 σ^2 的容许估计，我们作变换 $u=y/2$，于是 u 的分布对 $d\mu$ 的密度函数为

$$p(u;\sigma^2)\mathrm{d}u=\frac{1}{\sigma^n\Gamma(n/2)}u^{\frac{n}{2}-1}\mathrm{e}^{-u/\sigma^2}\mathrm{d}\mu(2u),$$

$$0\leqslant u<\infty,0<\sigma^2<\infty$$

它是 $\theta=\sigma^2,\alpha=n/2$ 的 Gamma 分布，所以

$$\frac{u}{1+\alpha}=\frac{y/2}{1+\alpha}=\frac{\frac{1}{2}\sum_{i=1}^{n}x_i^2}{1+\alpha}=\frac{1}{n+2}\sum_{i=1}^{n}x_i^2$$

是 $\theta=\sigma^2$ 的容许估计. 假如损失函数取

$$L(\sigma^2,d)=(d-\sigma^2)^2/\sigma^4$$

那么 $(n+2)^{-1}\sum_{i=1}^{n}x_i^2$ 仍是 σ^2 的容许估计.

§5.3.4 最小最大估计的容许性

下面两个定理表明了最小最大估计与容许性的关系.

定理 5.2 在一个统计决策问题中，假如 $\delta_0(x)$ 是参数 θ 的唯一最小最大估计，则 $\delta_0(x)$ 也是参数 θ 的容许估计.

证明：倘若 $\delta_0(x)$ 是非容许的，则应存在另一个估计 $\delta_1(x)\neq\delta_0(x)$，$a.s.[\mu]$，使得

$$R(\theta,\delta_1)\leqslant R(\theta,\delta_0),\forall\ \theta\in\Theta$$

而对某些 $\theta\in\Theta$，有严格不等式成立. 因此

$$\sup_{\theta\in\Theta}R(\theta,\delta_1)\leqslant\sup_{\theta\in\Theta}R(\theta,\delta_0)$$

从而 $\delta_1(x)$ 也是 θ 的最小最大估计，这与 δ_0 是唯一的最小最大估计矛盾，故此种 $\delta_1(x)$ 不存在，证毕.

从证明过程可以看出，定理 5.2 对任一指定的决策函数(估计)类都成立. 故没有指出具体的决策函数类. 下面的定理也是这样.

定理 5.3 在一个统计决策问题中，假如 $\delta_0(x)$ 是参数 θ 的容许估计，且在参数空间 Θ 上有常数风险，则 $\delta_0(x)$ 也是 θ 的最小最大估计.

证明：倘若 $\delta_0(x)$ 不是 θ 的最小最大估计，则存在另一个估计 $\delta_1(x)$，使得

$$\sup_{\theta\in\Theta}R(\theta,\delta_1)<\sup_{\theta\in\Theta}R(\theta,\delta_0)=R(\theta,\delta_0)=\text{Const.}$$

从而有

$$R(\theta,\delta_1) < R(\theta,\delta_0), \forall\ \theta \in \Theta$$

这表明，$\delta_0(x)$是非容许估计，这一矛盾证明 $\delta_0(x)$是 θ 的最小最大估计. 证毕.

例 5.17　设随机变量 Y 服从 Gamma 分布 $Ga(\alpha,1/\theta)$，其中 $\alpha>0$ 已知，θ 未知. 利用定理 5.1，曾在例 5.16 中，在平方损失函数 $L(\theta,\delta)=(\delta-\theta)^2$ 下求得 θ 的容许估计 $\hat{\theta}(y)=y/(1+\alpha)$，其风险函数为

$$R(\theta,\hat{\theta}) = \frac{1}{\theta^{\alpha}\Gamma(\alpha)}\int_0^{\infty}\left(\frac{y}{1+\alpha}-\theta\right)^2 y^{\alpha-1}e^{-y/\theta}dy$$
$$= \frac{\theta^2}{\Gamma(\alpha)}\int_0^{\infty}\left(\frac{u}{1+\alpha}-1\right)^2 u^{\alpha-1}e^{-u}du = \frac{\theta^2}{1+\alpha}$$

这个风险仍依赖于 θ. 假如取如下加权损失函数

$$L(\theta,\delta) = (\delta-\theta)^2/\theta^2$$

那么 $\hat{\theta}$ 的风险为$(1+\alpha)^{-1}$. 它与 θ 无关. 由定理 5.3 知，$\hat{\theta}(y)$在上述加权损失函数下是最小最大估计.

从例 5.16 和这个例子可以看出，在加权损失函数 $L(\theta,\delta)=\lambda(\theta)\cdot(\delta-\theta)^2$ 下求证 θ 的某个估计的容许性时，与 $\lambda(\theta)$的形式无关，但在讨论该估计是否是最小最大估计时，与 $\lambda(\theta)$的形式有很大关系. 只在特殊形式下，才能是最小最大估计.

§ 5.4　Bayes 决策准则

§ 5.4.1　先验分布

在叙述贝叶斯风险准则时需要先验分布的概念.

设有一个参数统计结构$(\mathscr{X},\mathscr{B},\{p_\theta:\theta\in\Theta\})$，其中参数 θ 是一个未知量. 为了对参数 θ 作统计推断(或统计决策)，我们需要从总体抽取样本，并且愈多愈好. 因为样本含有未知参数的信息，并且是“最新鲜”的信息. 这是经典统计推断的主要依据. 可是我们周围还存在有一些非样本信息. 它也可以用于统计推断和统计决策. 这些非样本信息主要来源于经验和历史资料. 由于这些经验和历史资料大多存在于(获得样本

的)试验之前,故又称为先验信息.

例 5.18 英国统计学家 Savage. L. J. 曾考虑如下两个统计试验.

a. 一位常饮牛奶加茶的妇女声称,她能辨别先倒进杯子里的是茶还是牛奶.对此做了十次试验,她都正确地说出了.

b. 一位音乐家声称,他能从一页乐谱辨别出是海顿(Haydn)还是莫扎特(Mozart)的作品.在十次这样的试验中,也都辨别正确.

在这两个统计试验中,假如认为被试验者是在猜测,每次成功概率为 0.5,那么十次都猜中的概率为 $2^{-10}=0.0009766$.这是很小的概率,是不可能发生的,所以认为"每次成功概率为 0.5"的原假设应被拒绝.被试验者每次成功概率要比 0.5 大得多.这就不是猜测,而是他们的经验帮了他们的忙.可见经验(先验信息的一种)在推断中不可忽视.

例 5.19 "免检产品"是怎样决定的?某工厂的产品每天都要抽 n 件,从中获得不合格品率 θ 的估计.经过一段时间后,这就可以根据历史资料(先验信息的一种)对过去产品的不合格品率 θ 构造一个分布

$$P(\theta = i/n) = \pi_i, \quad i = 0,1,\cdots,n$$

这种对先验信息进行加工获得的分布今后称为先验分布.有了这种先验分布就可得到该厂产品的不合格品率 θ 的一个全面看法.如果这个分布的概率绝大部分集中在 $\theta=0$ 附近,那该产品可以认为是"信得过产品".假如以后的历次抽检结果与历史资料提供的先验分布是一致的,那就可以对它作出"免检产品"的决定,或者每月抽检一次就足够了,这就省去了大量的人力和物力.

从上面两个例子可以看出,我们应注意先验信息的收集,挖掘和加工,使它数量化,形成先验分布,参加到统计推断中来,以提高统计推断的质量.忽视先验信息的利用,有时是一种浪费,有时还会导出不合理的结论.在统计中利用先验信息就形成 Bayes 统计.

Bayes 统计起源于英国学者贝叶斯(Bayes T. R. 1702(?)—1761)死后于 1763 年发表的一篇论文"论有关机遇问题的求解".在此文中提出著名的 Bayes 公式和一种归纳推理的方法.之后被一些统计学家发展成一种系统的统计推断方法.到本世纪 30 年代已形成 Bayes 学派,到 50～60 年代已发展成一个有影响的统计学派.其影响还在日益

扩大.

Bayes 学派的最基本的观点是：任一未知量 θ 都可看作随机变量，可用一个概率分布去描述，这个分布称为先验分布. 因为任一未知量都有不确定性. 而在表述不确定性时，概率与概率分布是最好的语言. 例 5.19 中产品的不合格品率 θ 是未知的，但每天都在变化，把它看作随机变量是合理的，用一个概率分布去描述它是很恰当的.

关于未知量是否可看作随机变量在经典学派与 Bayes 学派间争论了很长时间. 如今经典学派已不反对这一观点. 著名的美国经典统计学家 Lehmann. E. L 在他的《点估计理论》一书中写道："把统计问题中的参数看作随机变量的实现要比看作未知参数更合理些."如今两派的争论焦点是：如何利用各种先验信息合理地确定先验分布，这在有些场合是容易解决的，但在很多场合还是相当困难的. 这时应加强研究，切不可简单处理. 草率地定下先验分布. 要知道，这是为指责 Bayes 方法提供炮弹.

关于先验分布的确定至今已有一些成功的方法，Berger 的书[2]的第三章对此作了较为全面的叙述. 本章内容将要讲述其中的一些方法. 这里我们先介绍用主观概率确定离散先验分布方法.

Bayes 学派认为一个事件的概率可以是人们根据经验对该事件发生可能性所给出的个人信念. 这样给出的概率常称为主观概率. 譬如：

一个企业家认为"一项新产品在未来市场上畅销的"概率是 0.8，这里的 0.8 是根据他自己多年经验和当时的一些市场信息综合而成的个人信念.

一位脑外科医生要对一位病人动手术，他认为成功的概率是 0.9，这是他根据手术的难易程度和自己的手术经验而对"手术成功"所给出的把握程度.

一位教师认为甲学生考取大学的概率是 0.95，而乙学生考取大学的概率是 0.5. 这是教师根据自己多年的教学经验和甲、乙两位学生的学习情况而分别给出的个人信念.

这样的例子在我们生活，生产和经济活动中也是常遇到的. 他们给出的主观概率决不是随意的，而是要求当事人对所考察的事件有较透

彻的了解和丰富的经验，甚至是这一行的专家，并能对周围信息和历史信息进行仔细分析，在这个基础上确定的主观概率就能符合实际. 所以应把主观概率与主观臆造，瞎说一通区别开来.

自主观概率提出以来，使用的人愈来愈多，特别在经济领域和决策分析中使用较为广泛. 因为在那里遇到的随机现象大多是不能重复的. 无法用频率方法去确定事件概率. 在这个意义上看，主观概率至少使人们在频率方法不适用时也可谈论概率，使用统计方法.

下面通过一些例子来说明确定主观概率的方法.

例 5.20 一位出版商要知道一本新书能畅销（事件 A）的概率是多少，以便决定是否与作者签订出版合同. 他在了解这本新书的内容后，根据他自己多年出书的经验认为：该书畅销的可能性较大，畅销（A）比不畅销（$\overline{A}$）的可能性要高出一倍，即 $P(A)=2P(\overline{A})$. 由此根据概率性质 $P(A)+P(\overline{A})=1$ 可以推得 $P(A)=2/3$，即该书畅销的主观概率是 2/3.

这种用对立事件的比较来确定主观概率是最简单方法. 假如对 A 与 $\overline{A}$ 说不出哪一个发生的可能性较大，那就定 $P(A)=1/2$. 这个方法可以推广到多个事件上. 这种方法虽不严格，但符合人们在实际中看待概率的直观方式.

例 5.21 有一项带有风险的生意，欲估计成功（记为 A）的概率. 为此决策者去拜访这方面的专家（如董事长，银行家等），向专家提这样的问题："如果这种生意做 100 次，你认为能成功几次？"专家回答："成功次数不会太多，大约 60 次."这时 $P(A)=0.6$ 是专家的主观概率，可此专家还不是决策者. 决策者很熟悉这位专家，认为他的估计往往是偏保守的，过分谨慎的. 决策者决定修改专家的估计，把 0.6 提高到 0.7. 这样 $P(A)=0.7$ 就是决策者的主观概率.

这种用专家意见来确定主观概率的方法是常用的. 当决策者对某事件了解甚少时，就可去征求专家意见. 这里要注意两点. 一是向专家提出的问题要设计好，既要使专家易懂，又要使专家回答不是模棱两可；二是要对专家本人较为了解，以便作出修正，形成决策者自己的主观概率.

例 5.22　某公司经营儿童玩具多年，今设计了一种新式玩具将投入市场. 现要估计此新式玩具在未来市场上的销售状况. 经理查阅了本公司过去 37 种新式玩具的销售记录，得知销售状态是畅销(θ_1)，一般(θ_2)，滞销(θ_3)分别有 29，6，2 种，于是算得过去新式玩具的三种销售状态的概率分别为

$$\frac{29}{37}=0.784,\quad \frac{6}{37}=0.162,\quad \frac{2}{37}=0.054$$

考虑到这次设计的新玩具不仅外形新颖，而且在开发儿童智力上有显著突破，经理认为此种新玩具会更畅销一些，滞销可能性更小，故对上述概率作了修改，提出自己的主观概率如下：

$$P(\theta_1)=0.85,\quad P(\theta_2)=0.14,\quad P(\theta_3)=0.01$$

这个例子表明，假如有历史数据，要尽量利用，帮助形成初步概念，然后作一些对比，修正，再形成个人信念，这对给出主观概率大有好处.

依据经验和历史资料等先验信息给出的主观概率没有什么固定的模式，但所确定的主观概率都必须满足概率的三条公理，发现与三条公理及其性质有不和谐之处，必须立即修正，直到和谐为止. 这时给出的主观概率才能称得上概率.

§5.4.2　Bayes 风险准则

这里和以后，除非特别声明，我们总把讨论限制在非随机化决策函数类 $\mathscr{D}$ 中.

定义 5.9　在给定的统计决策问题中，设 $R(\theta,\delta(x))$ 为决策函数 $\delta(x)$ 的风险函数，$\pi(\theta)$ 为 θ 的先验分布，则平均风险

$$R_\pi(\delta)=\begin{cases}\int_\Theta R(\theta,\delta)\pi(\theta)\mathrm{d}\theta, & \theta\text{ 为连续}\\ \sum_i R(\theta_i,\delta)\pi(\theta_i), & \theta\text{ 为离散}\end{cases}\tag{5.31}$$

称为 $\delta(x)$ 的 Bayes 风险. 假如在决策函数类 $\mathscr{D}$ 中存在这样的决策函数 $\delta^*(x)$，使得

$$R_\pi(\delta^*)=\inf_{\delta\in\mathscr{D}}R_\pi(\delta)\tag{5.32}$$

则称 $\delta^*(x)$ 为决策函数类 $\mathscr{D}$ 在 Bayes 风险准则下的最优决策函数，简

称 Bayes 决策函数或 Bayes 解. 在估计问题中，它就称为 Bayes 估计.

从上述定义可见，Bayes 风险是一个实数，它是按先验分布算出的平均风险，$\mathscr{D}$ 中每个决策函数 $\delta(x)$ 都有一个平均风险，Bayes 风险准则就是用平均风险来评价其优劣. 平均风险愈小愈好，$\mathscr{D}$ 中具有最小平均风险的决策函数便是 Bayes 风险决策函数. 显然，上述评价是对固定的先验分布 $\pi(\theta)$ 而言的. 当先验分布改变时，Bayes 决策函数也会改变的.

例 5.23 对例 5.3 的二行动线性决策问题在例 5.8 中分别给出 8 个决策函数的风险函数. 根据该厂的近一年的历史资料整理，该厂不合格品率 θ 不会超过 0.12，而位于区间(0.04，0.08)内占 80%，而低于 0.04 或高于 0.08 的各占 10%，这样一来我们就得到如图 5.5 所示的先验分布 $\pi(\theta)$.

$$\pi(\theta)=\begin{cases}2.5, & 0<\theta\leqslant 0.04\\ 20.0, & 0.04<\theta\leqslant 0.08\\ 2.5, & 0.08<\theta\leqslant 0.12\\ 0, & \text{其它}\end{cases}$$

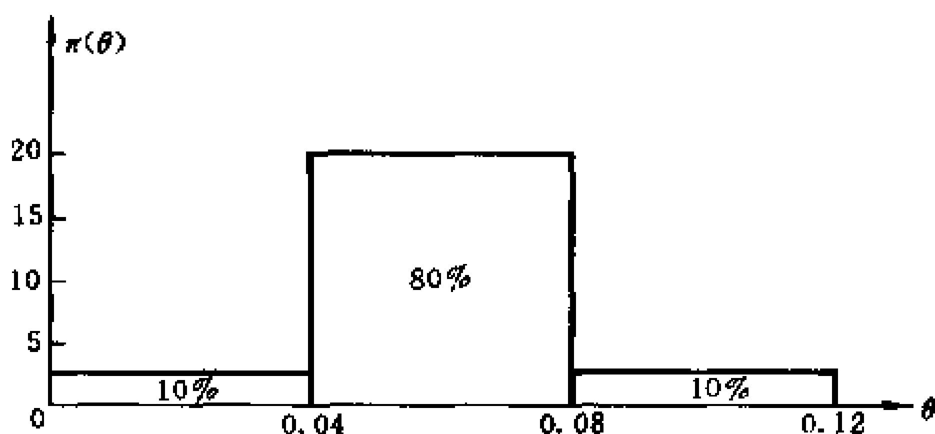


图 5.5 不合格品率 θ 的先验分布

利用这个先验分布我们可在这 8 个决策函数中寻找 Bayes 决策函数. 为此我们要利用例 5.8 中给出 8 个风险函数的表达式，逐一计算它们对此先验分布的平均风险. 譬如，$\delta_1(x)$ 和 $\delta_2(x)$ 的平均风险分别为

$$R_\pi(\delta_1) = \int_0^{\theta_0} (78.4 - 1250\theta)\pi(\theta)\mathrm{d}\theta = 11.7925$$

$$R_\pi(\delta_2) = \int_0^{\theta_0} (78.4 - 1250\theta - 78.4\theta^2 + 1250\theta^3)\pi(\theta)\mathrm{d}\theta$$

$$+ \int_{\theta_0}^{0.12} (-78.4\theta^2 + 1250\theta^3)\pi(\theta)\mathrm{d}\theta = 11.8467$$

其中 $\theta_0 = 0.062\,72$，类似地可计算另外 6 个平均风险，可得

$$R_\pi(\delta_3) = 12.4092,\quad R_\pi(\delta_4) = 12.4634$$

$$R_\pi(\delta_5) = 7.7215,\quad R_\pi(\delta_6) = 7.7758$$

$$R_\pi(\delta_7) = 8.3382,\quad R_\pi(\delta_8) = 8.3925$$

比较这 8 个平均风险，可以看出 $\delta_5(x)$ 的平均风险最小，故在统计决策问题中 $\delta_5(x)$ 是在此先验分布下 θ 的 Bayes 决策函数.

§ 5.4.3　Bayes 公式

Bayes 公式的离散形式已为大家熟知，这里用密度函数再一次叙述 Bayes 公式，从中介绍 Bayes 学派的一些观点.

1. 依赖于参数 θ 的密度函数在经典统计中记为 $p(x;\theta)$ 或 $p_\theta(x)$. 它表示 Θ 中不同的 θ 对应不同的分布. Bayes 学派认为该密度函数是在随机变量 θ 给定某个值时，X 的条件密度函数，故应记为 $p(x|\theta)$.

2. 根据参数 θ 的先验信息确定先验分布 $\pi(\theta)$.

3. 从 Bayes 观点看，样本 $X = (X_1, \cdots, X_n)$ 的产生要分两步进行. 首先设想从先验分布 $\pi(\theta)$ 产生一个观察值 θ，然后再从条件分布 $p(x|\theta)$ 产生样本观察值 $x = (x_1, \cdots, x_n)$. 这时样本 X 的联合条件密度函数为

$$p(x|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

这个联合分布综合了样本信息，又称为似然函数.

4. 由于 θ 是设想出来的，仍然是未知的，欲要把先验信息与样本信息综合，不能仅用设想值，而应用 $\pi(\theta)$. 这样一来，样本 X 与参数 θ 的联合分布

$$h(x,\theta) = p(x|\theta)\pi(\theta) \tag{5.33}$$

把先验信息与样本信息都综合到一起了.

5. 我们的任务是要对未知参数 θ 作出统计决策，为此把联合分布进行如下分解

$$h(x,\theta) = \pi(\theta|x)m(x)$$

其中 $m(x)$ 是 x 的边际密度函数

$$m(x) = \int_{\Theta} h(x,\theta)\mathrm{d}\theta = \int_{\Theta} p(x|\theta)\pi(\theta)\mathrm{d}\theta \tag{5.34}$$

它与 θ 无关，或者说 $m(x)$ 不含 θ 的任何先验信息. 因此能用来对 θ 作出统计决策的仅是条件分布 $\pi(\theta|x)$. 它的计算公式是

$$\pi(\theta|x) = \frac{h(x,\theta)}{m(x)} = \frac{p(x|\theta)\pi(\theta)}{\int_{\Theta} p(x|\theta)\pi(\theta)\mathrm{d}\theta} \tag{5.35}$$

这就是 Bayes 公式的密度函数形式. 这个在样本给定下 θ 的条件分布称为 θ 的后验分布. 它是在样本给定下集中了样本与先验中有关 θ 的一切信息. 它要比先验分布 $\pi(\theta)$ 更接近于实际情况. 因此使用后验分布 $\pi(\theta|x)$ 对 θ 作统计决策会得到改进.

(5.35)式是在 X 和 θ 都是连续随机变量场合下的 Bayes 公式. 其它场合下的 Bayes 公式亦可类似写出. 譬如，当 X 是离散随机变量时，只要把(5.35)中的密度函数 $p(x|\theta)$ 改为概率 $P(X=x|\theta)$ 即可. 而当 θ 是离散随机变量时，只要把(5.35)中的密度函数 $\pi(\theta)$ 改为分布列 $\pi(\theta_i)$，$i=1,2,\cdots$，积分号改为求和号即可.

6. 回忆上述过程，不难看出，当从总体获得样本 X 后，Bayes 公式把人们对 θ 的认识由 $\pi(\theta)$ 调正到 $\pi(\theta|x)$ 这个调整过程可形象地表示为

$$\text{先验信息} \oplus \text{样本信息} \Rightarrow \text{后验信息}$$
$$\pi(\theta) \oplus p(x|\theta) \Rightarrow \pi(\theta|x)$$

其中符号“$\oplus$”应理解为 Bayes 公式的作用. 下面的例子可以更具体地说明这个过程.

例 5.24　设事件 A 出现的概率为 θ，即 $P(A)=\theta$，为了估计 θ 而做了 n 次独立观察，其中事件 A 出现次数 X 服从二项分布 $b(n,\theta)$，即

$$P(X=x\mid\theta)=\binom{n}{x}\theta^{x}(1-\theta)^{n-x},\quad x=0,1,\cdots,n$$

这就是似然函数.假如在试验之前我们对事件 A 没有什么了解,从而对其发生的概率 θ 也说不出是大是小.在这种场合,Bayes 建议用区间(0,1)上的均匀分布 $U(0,1)$ 作为 θ 的先验分布.因为它在(0,1)上每一点上都机会均等.Bayes 的这个建议被后人称为 Bayes 假设.这时 θ 的先验分布为 $\pi(\theta)=U(0,1)$.为了综合试验信息和先验信息,可以利用 Bayes 公式.为此先计算样本 X 与参数 θ 的联合分布

$$h(x,\theta)=\binom{n}{x}\theta^{x}(1-\theta)^{n-x},\quad x=0,1,\cdots,n,0<\theta<1$$

从形式上看上述联合分布与二项分布没有差别,可在定义域上有差别.再计算样本 X 的边际分布

$$m(x)=\int_0^1 h(x,\theta)\mathrm{d}\theta=\binom{n}{x}\int_0^1\theta^{x}(1-\theta)^{n-x}\mathrm{d}\theta$$

$$=\binom{n}{x}\frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)}$$

由此得 θ 的后验分布

$$\pi(\theta\mid x)=\frac{h(x,\theta)}{m(x)}=\frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)}\cdot\theta^{(x+1)-1}(1-\theta)^{(n-x+1)-1},\quad 0<\theta<1$$

这就是参数为 $x+1$ 和 $n-x+1$ 的 Beta 分布 $Be(x+1,n-x+1)$.

进一步研究这个例子.考察样本信息 x 是如何对先验分布作调整的.试验前,θ 在(0,1)上均匀分布(见图 5.6 左侧).当试验结果 $X=x$ 时,θ 的后验分布 $\pi(\theta|x)$ 虽仍在(0,1)上取值,但已不是均匀分布,而是一个单峰分布,其峰的位置是随着 x 的增大而逐渐向右移(见图 5.6 的右侧).不论哪种情况发生,其峰值总在 x/n 处达到.譬如在 $X=0$ 时,它表示在 n 次试验中事件 A 一次也没出现,这表明事件 A 发生的概率很小.当它与均匀分布结合后所得后验分布反映了这个信息,θ 在 0 附近取值的可能性很大,在 1 附近取值的可能性很小.类似地,在 $X=n$ 时,所得的后验分布在 1 附近取值的可能性很大,而在 0 附近取值的可

能性小. 当 $x<n/2$ 时, $x/n<1/2$, 后验分布的峰偏左; 当 $x>n/2$ 时, 后验分布的峰偏右; 当 $x=n/2$(n 为偶数), 后验分布对称, 其峰在 1/2 处.

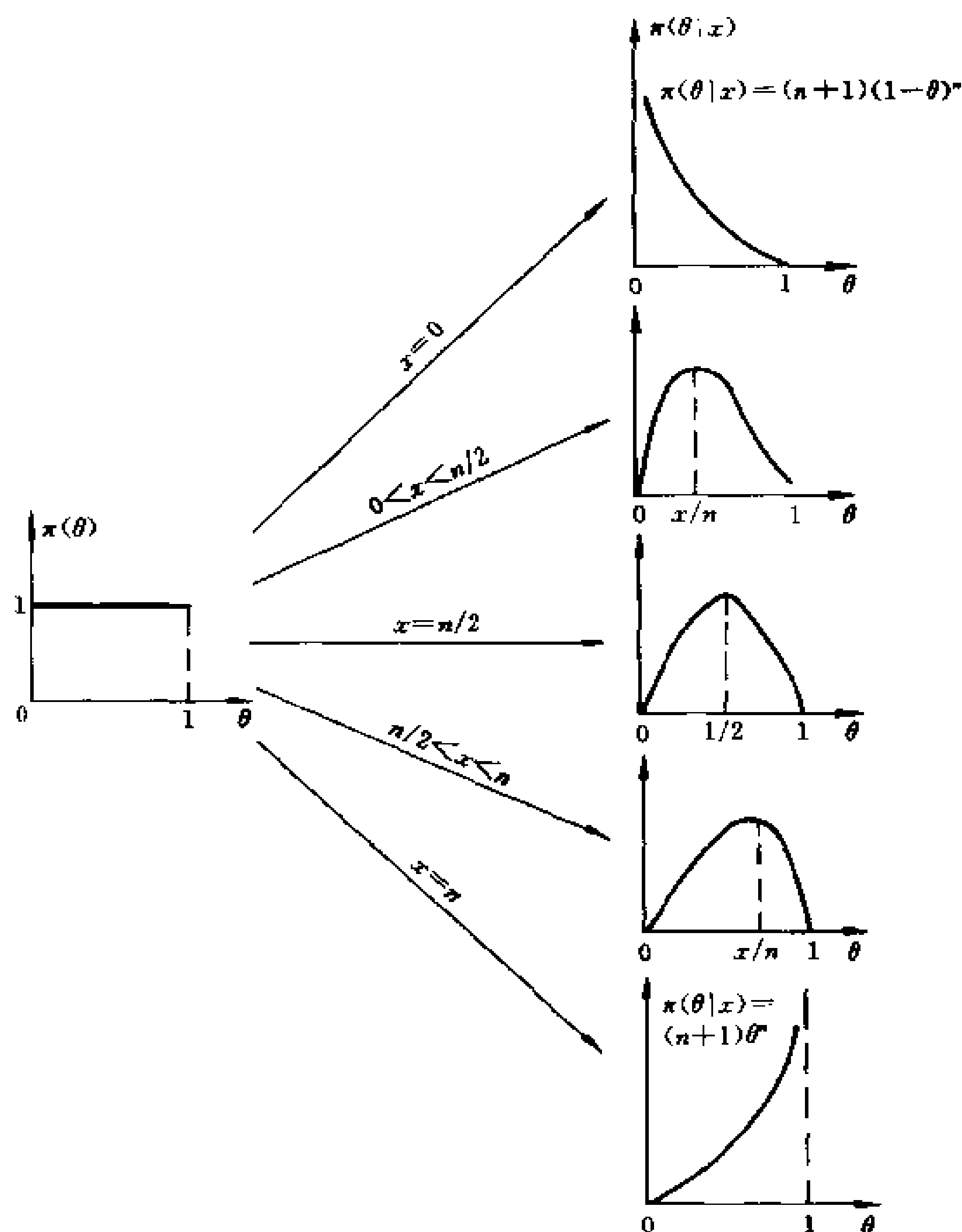


图 5.6 试验信息对先验分布的调整

7. 最后指出 Bayes 公式的一个性质.

定理 5.4 设有一个可控参数结构$(\mathscr{X},\mathscr{B},\{p_\theta:\theta\in\Theta\})$，$\mu$是其控制测度. 又设$T=T(x)$是$(\mathscr{X},\mathscr{B})$到$(\mathscr{T},\mathscr{C})$上的一个充分统计量. $p^T(t|\theta)$是T关于诱导测度μ^T的密度函数. 则由$p^T(t|\theta)$导出的后验分布$\pi^T(\theta|t)$与用样本密度函数$p(x|\theta)$导出的后验分布$\pi(\theta|x)$是几乎处处相等，即

$$\pi^T(\theta|t)=\pi(\theta|x)\qquad \text{a.s.}\ [\mu] \tag{5.36}$$

证明：由因子分解定理知，存在一个非负$\mathscr{B}$可测函数$h(x)$和一个非负$\mathscr{C}$可测函数$g(t|\theta)$，使得

$$p(x|\theta)=g(t(x)|\theta)h(x)\qquad \text{a.s.}\ [\mu]$$

因此在给定x下，θ的后验密度为

$$\pi(\theta|x)=\frac{g(t(x)|\theta)\pi(\theta)}{\int_\Theta g(t(x)|\theta)\pi(\theta)\nu(\mathrm{d}\theta)}$$

其中$\nu(\mathrm{d}\theta)$是$\pi(\theta)$的控制测度. 另外，对任意$c\in\mathscr{C}$，用T的密度函数和X的密度函数可分别算得

$$P_\theta^T(C)=\int_C p^T(t|\theta)\mu^T(\mathrm{d}t)$$

$$P_\theta[T^{-1}(C)]=\int_{T^{-1}(C)}p(x|\theta)\mu(\mathrm{d}x)$$

由于$P_\theta^T(C)=P_\theta[T^{-1}(C)]$，故对任意的$C\in\mathscr{C}$，有

$$\int_C p^T(t|\theta)\mu^T(\mathrm{d}t)=\int_{T^{-1}(C)}p(x|\theta)\mu(\mathrm{d}x)$$

由此可知，$p^T(t|\theta)$是$p(x|\theta)$在条件$T=t$下的条件期望，即

$$\begin{aligned}p^T(t|\theta)&=\mathrm{E}[p(x|\theta)|T=t]\\&=\mathrm{E}[g(t|\theta)h(x)|T=t]\\&=g(t|\theta)\mathrm{E}[h(x)|T=t]\\&=g(t|\theta)J(t)\qquad \text{a.s.}\ [\mu^T]\end{aligned}$$

其中$J(t)=\mathrm{E}[h(x)|T=t]$是非负函数. 于是在$T=t$下，$\theta$的后验密度函数为

$$\pi^T(\theta|t)=\frac{p^T(t|\theta)\pi(\theta)}{\int_\Theta p^T(t|\theta)\pi(\theta)\nu(\mathrm{d}\theta)}$$

$$= \frac{g(t|\theta)\pi(\theta)}{\int_{\Theta} g(t|\theta)\pi(\theta)\nu(\mathrm{d}\theta)} \qquad \text{a.s.}\,[\mu^T]$$

比较上述二个后验密度，可得(5.36)．证毕．

这个定理表明：在有充分统计量的场合，后验分布可通过充分统计量的分布求得．

例 5.25 设 $X=(X_1,\cdots,X_n)$是来自正态总体 $N(\xi,\sigma^2)$的一个样本，样本的联合密度函数为

$$p(x|\xi,\sigma)=(2\pi)^{-\frac{n}{2}}\sigma^{-n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\xi)^2\right\}$$

$$=(2\pi)^{-\frac{n}{2}}\sigma^{-n}\exp\left\{-\frac{Q}{2\sigma^2}-\frac{n}{2\sigma^2}(\overline{x}-\xi)^2\right\}$$

其中 $\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，$Q=\sum_{i=1}^{n}(x_i-\overline{x})^2$ 是正态分布族的充分统计量的取值．

若设参数 $\theta=(\xi,\sigma)$的联合先验密度为 $\pi(\xi,\sigma)$，则在样本 $x=(x_1,\cdots,x_n)'$下，$\theta=(\xi,\sigma)$的后验密度为

$$\pi(\xi,\sigma|x)=\frac{\sigma^{-n}\pi(\xi,\sigma)\exp\left\{-\frac{Q}{2\sigma^2}-\frac{n}{2\sigma^2}(\overline{x}-\xi)^2\right\}}{\int_{-\infty}^{\infty}\left\{\int_0^{\infty}\sigma^{-n}\pi(\xi,\sigma)\exp\left\{-\frac{Q}{2\sigma^2}-\frac{n}{2\sigma^2}(\overline{x}-\xi)^2\right\}\mathrm{d}\sigma\right\}\mathrm{d}\xi}$$

另一方面，在正态分布族中，$\overline{X}$ 与 Q 是两个相互独立的统计量，且 $\overline{X}\sim N(\xi,\sigma^2/n)$，$Q/\sigma^2\sim\chi^2(n-1)$．所以充分统计量 $T=(\overline{X},Q)$的联合密度函数为

$$p^T(\overline{x},Q|\xi,\sigma)=p(\overline{x}|\xi,\sigma)p(Q|\sigma)$$

$$=k\sigma^{-1}\exp\left\{-\frac{n}{2\sigma^2}(\overline{x}-\xi)^2\right\}(\sigma^2)^{-\frac{n-1}{2}}\exp\left\{-\frac{Q}{2\sigma^2}\right\}$$

其中 k 是与 ξ,σ 无关的常量．在给定 $T=(\overline{X},Q)$的条件下，此密度函数与先验密度 $\pi(\xi,\sigma)$结合而产生的后验密度为

$$\pi^T(\xi,\sigma|\overline{x},Q)=\frac{p^T(\overline{x},Q|\xi,\sigma)\pi(\xi,\sigma)}{\int_{\Theta}p^T(\overline{x},Q|\xi,\sigma)\pi(\xi,\sigma)\mathrm{d}\xi\mathrm{d}\sigma}$$

$$= \frac{\sigma^{-n}\pi(\xi,\sigma)\exp\left\{-\frac{Q}{2\sigma^2}-\frac{n}{2\sigma^2}(\overline{x}-\xi)^2\right\}}{\int_{-\infty}^{\infty}\left\{\int_0^{\infty}\sigma^{-n}\pi(\xi,\sigma)\exp\left\{-\frac{Q}{2\sigma^2}-\frac{n}{2\sigma^2}(\overline{x}-\xi)^2\right\}\mathrm{d}\sigma\right\}\mathrm{d}\xi}$$

比较这两个后验密度，可得

$$\pi^T(\xi,\sigma|\overline{x},Q) = \pi(\xi,\sigma|x)$$

§5.4.4　共轭先验分布

大家知道，在区间(0,1)上均匀分布是 Beta 分布 $Be(1,1)$，这样从例 5.24 中可以看到一个有趣的现象。二项分布 $b(n,\theta)$中的成功概率 θ 的先验分布若取 $Be(1,1)$，则其后验分布也是 Beta 分布 $Be(x+1,n-x+1)$。先验分布与后验分布同属于一个 Beta 分布族，只不过参数不同罢了。这一现象不是偶然的，假如把 θ 的先验分布换成一般的 Beta 分布 $Be(a,b)$，其中 $a>0,b>0$。则经过类似计算可以看出 θ 的后验分布仍然是 Beta 分布 $Be(a+x,b+n-x)$，此种先验分布称为 θ 的共轭先验分布。在其它场合还会遇到其它共轭先验分布，它的一般定义如下：

定义 5.10　设 $\mathscr{P}=\{p(x|\theta):\theta\in\Theta\}$是以 θ 为参数的密度函数族，又设 $\mathscr{H}=\{\pi(\theta)\}$是 θ 的先验分布族。假如对任意的 $p\in\mathscr{P}$ 和 $\pi\in\mathscr{H}$，所得的后验分布 $\pi(\theta|x)$仍在族 $\mathscr{H}$ 中，则称 $\mathscr{H}$ 为 $\mathscr{P}$ 的共轭分布族，或者称 $\pi(\theta)$为参数 θ 的共轭先验分布。

从这个定义可以看出，共轭先验分布是对某一分布中的参数而言的。离开指定参数及其所在的分布，谈论共轭先验分布是没有意义的。常用的共轭先验分布列于表 5.4 中。

表 5.4　常用的共轭先验分布

总体分布	参数	共轭先验分布
二项分布	成功概率	Beta 分布
Poisson 分布	均值	Gamma 分布
指数分布	均值倒数	Gamma 分布
正态分布(方差已知)	均值	正态分布
正态分布(均值已知)	方差	倒 Gamma 分布

注：若 $X\sim Ga(a,\lambda)$，则 X^{-1}的分布称为倒 Gamma 分布

例 5.26 正态均值(方差已知)的共轭先验分布是正态分布.设 $X_1, X_2, \cdots, X_n$,是来自正态分布 $N(\theta, \sigma^2)$ 的一个样本.其中 σ^2 已知,此样本的联合密度函数为

$$p(x|\theta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2\right\},$$

$$-\infty < x_1, \cdots, x_n < \infty \tag{5.37}$$

现取另一正态分布 $N(\mu, \tau^2)$ 作为正态均值 θ 的先验分布,即

$$\pi(\theta) = \frac{1}{\sqrt{2\pi}\tau}\exp\left\{-\frac{1}{2\tau^2}(\theta - \mu)^2\right\},$$

$$-\infty < \theta < \infty \tag{5.38}$$

其中 μ 与 τ^2 为已知.由此可写出样本 x 与参数 θ 的联合密度函数

$$h(x, \theta) = K_1 \exp\left\{-\frac{1}{2}\left[\frac{n\theta^2 - 2n\theta\bar{x} + \sum_{i=1}^n x_i^2}{\sigma^2} + \frac{\theta^2 - 2\mu\theta + \mu^2}{\tau^2}\right]\right\}$$

其中 $K_1 = (2\pi)^{-(n+1)/2}\tau^{-1}\sigma^{-n}$, $\bar{x} = \sum_{i=1}^n x_i/n$, 若再记

$$\sigma_0^2 = \frac{\sigma^2}{n}, A = \frac{1}{\sigma_0^2} + \frac{1}{\tau^2}, B = \frac{\bar{x}}{\sigma_0^2} + \frac{\mu}{\tau^2}, C = \frac{\sum_{i=1}^n x_i^2}{\sigma^2} + \frac{\mu^2}{\tau^2}$$

则有

$$h(x, \theta) = K_1 \exp\left\{-\frac{1}{2}[A\theta^2 - 2\theta B + C]\right\}$$

$$= K_1 \exp\left\{-\frac{(\theta - B/A)^2}{2/A} - \frac{1}{2}(C - B^2/A)\right\}$$

由此容易算得样本 X 的边际分布

$$m(x) = \int_{-\infty}^{\infty} h(x, \theta) d\theta = K_1 \exp\left\{-\frac{1}{2}(C - B^2/A)\right\}\left(\frac{2\pi}{A}\right)^{\frac{1}{2}}$$

上两式相除,即得 θ 的后验分布

$$\pi(\theta|x) = \left(\frac{2\pi}{A}\right)^{-\frac{1}{2}} \exp\left\{-\frac{(\theta - B/A)^2}{2/A}\right\} \tag{5.39}$$

这是正态分布,其均值 μ_1 与方差 σ_1^2 分别为

$$\mu_1 = \frac{B}{A} = \frac{\bar{x}\sigma_0^{-2} + \mu\tau^{-2}}{\sigma_0^{-2} + \tau^{-2}}, \quad \sigma_1^2 = (\sigma_0^{-2} + \tau^{-2})^{-1} \tag{5.40}$$

譬如,数 $X\sim N(\theta,2^2)$,$\theta\sim N(10,3^2)$. 若从总体 X 中抽得 5 个样本,算得 $\bar{x}=12.1$. 于是可从(5.40)式算得 $\mu_1=11.93$,$\sigma_1^2=(6/7)^2$. 这时 θ 的后验分布为正态分布 $N\left(11.93,\left(\frac{6}{7}\right)^2\right)$. 由此可见,正态均值(方差已知)的共轭先验分布是正态分布.

另外,从(5.40)式中可以看出,后验均值 μ_1 是先验均值 μ 和样本均值 $\bar{x}$ 的加权之和,即

$$\mu_1=\lambda\mu+(1-\lambda)\bar{x}$$

其中权 $\lambda=\tau^{-2}/(\sigma_0^{-2}+\tau^{-2})$. 且先验方差 τ^2 愈小,先验均值在后验均值中占的比重愈大. 另一方面,当样本量 n 愈大时,$\sigma_0^2=\sigma^2/n$ 愈小,从而样本均值 $\bar{x}$ 在后验均值中占的比重也愈大,特别当 n 无限增大时,先验均值在后验均值中已是微不足道. 这种现象(从参数的变化看先验和样本的作用)常在使用共轭先验分布中见到.

从上述例子的计算中还可看到一个有趣的现象. 正态分布之所以能成为正态均值 θ 的共轭先验分布,关键在于样本分布(5.37)与先验分布(5.38)的密度函数都是 θ 的指数函数,并且指数上又都是 θ 的二次函数,故它们相乘,指数相加后,仍是同类型的指数函数,这就导致共轭性. 在统计学中常称此指数函数为正态分布的核. 譬如,若 $X\sim N(\mu,\sigma^2)$,其核为 $\exp\{-(x-\mu)^2/2\sigma^2\}$. 这时正态分布的密度函数 $p(x)$ 可表示为

$$p(x)\propto\exp\{-(x-\mu)^2/2\sigma^2\},\quad -\infty<x<\infty$$

其中符号"$\propto$"表示"正比于"之意,即符号 $\propto$ 两侧只相差一个不依赖于 x 的因子. 利用分布的正则性,不难恢复这个因子. 譬如,若 $p(x)\propto\exp\{-(x-5)^2\}$,可以看出 $X\sim N(5,1/2)$,于是可写出其密度函数 $p(x)=\frac{1}{\sqrt{\pi}}\exp\{-(x-5)^2\}$.

分布的核不仅正态分布有,其它分布也有. 常用分布的核有如下几种:

(1) 二项分布 $b(n,\theta)$ 的核是 $\theta^x(1-\theta)^{n-x}$.

(2) Poisson 分布 $P(\lambda)$ 的核是 $\lambda^x e^{-\lambda}$.

(3) Beta 分布 $Be(a,b)$的核是 $x^{a-1}(1-x)^{b-1}$.

(4) Gamma 分布 $Ga(\alpha,\lambda)$的核是 $x^{\alpha-1}e^{-\lambda x}$.

(5) 正态分布 $N(\mu,\sigma^2)$的核是 $\exp\{-(x-\mu)^2/2\sigma^2\}$.

熟悉分布的核可以简化后验分布的计算. 后验分布 $\pi(\theta|x)$的核一定是 θ 的某个函数,故用来计算 $\pi(\theta|x)$的 Bayes 公式的分母 $m(x)$是与 θ 无关的因子,可以略去,从而有

$$\pi(\theta|x) \propto p(x|\theta)\pi(\theta) \tag{5.41}$$

假如再用 $p(x|\theta)$与 $\pi(\theta)$的核代入,上式还可以简化,譬如在例 5.26 中用 $p(x|\theta)$和 $\pi(\theta)$的核去代替它们自己,则有

$$\begin{aligned}\pi(\theta|x) &\propto \exp\left\{-\frac{1}{2}\left[\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{\sigma^2}+\frac{(\theta-\mu)^2}{\tau^2}\right]\right\}\\ &\propto \exp\left\{-\frac{1}{2}[A\theta^2-2B\theta]\right\}\\ &\propto \exp\left\{-\frac{A}{2}(\theta-B/A)^2\right\}\end{aligned}$$

其中 A 与 B 如前所述,由此可立即可看出此后验分布定为正态分布 $N(B/A,1/A)$. 这就简化了寻求后验分布的计算过程.

例 5.27　二项分布中成功概率 θ 的共轭先验分布是 Beta 分布. 设总体 $X\sim b(n,\theta)$,其核为 $\theta^x(1-\theta)^{n-x}$,若取 Beta 分布 $Be(a,b)$,作为 θ 的先验分布,其核为 $\theta^{a-1}(1-\theta)^{b-1}$,从而可得 θ 的后验分布的核为

$$\pi(\theta|x) \propto \theta^{a+x-1}(1-\theta)^{b+n-x-1}$$

这就是 Beta 分布 $Be(a+x,b+n-x)$的核. 故此后验分布为

$$\pi(\theta|x)=\frac{\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(b+n-x)}\theta^{a+x-1}(1-\theta)^{b+n-x-1},$$

$$0<\theta<1 \tag{5.42}$$

可见,Beta 分布是成功概率 θ 的共轭先验分布.

另外,从 Beta 分布的均值可以看出后验分布在(5.42)的均值亦可表示为先验均值 $E(\theta)=a/(a+b)$和样本均值 $\bar{x}=x/n$ 的加权和,即

$$\frac{a+x}{a+b+n}=\frac{a}{a+b}\frac{a+b}{a+b+n}+\frac{x}{n}\frac{n}{a+b+n}$$

其中，权 $\lambda=\frac{a+b}{a+b+n}$. 当样本量 n 无限增大时，先验均值在后验均值中已是微不足道.

共轭先验分布在很多场合被采用. 其主要原因是计算极为简单，一般无需直接计算边际分布 $m(x)$，而后验分布 $\pi(\theta|x)$ 可从其 θ 的因式（核）得到. 在其它场合，$m(x)$ 与 $\pi(\theta|x)$ 的计算都会遇到困难. 譬如当 $X\sim N(\theta,\sigma^2)$ 和 θ 服从Cauchy分布时，则其后验分布 $\pi(\theta|x)$ 只能作数值计算，但是我们在选用共轭先验分布时要注意先验的合理性. 因为计算上的方便与先验的合理性相比那还是第二位的. 何况近十多年来统计计算技术得到长足的发展（见本书第六章），后验分布及各种后验量的计算已不是使用Bayes分析的障碍，这时更应注意选用先验分布的合理性.

先验分布中所含的参数（假如有的话）称为超参数. 共轭先验分布中常含有超参数，如何确定这些超参数是选用共轭先验分布遇到的另一个问题. 这个问题的回答没有一般程式，要看你手中握有的先验信息（经验和历史资料）的情况而定. 下面的例子可启发我们，确定超参数有几条途径可供选择.

例 5.28 前面已指出：二项分布中成功概率 θ 的共轭先验分布是Beta分布 $Be(a,b)$，现在来讨论此共轭先验分布中的二个超参数 a 与 b 如何确定. 下面分几种情况讨论：

a) 假如根据先验信息能获得成功概率的若干个（间接）观察值：$\theta_1,\theta_2,\cdots,\theta_n$. 一般它们是从历史数据整理加工获得的，由此可算得先验均值 $\bar{\theta}$ 和先验方差 S_θ^2 为：

$$\bar{\theta}=\frac{1}{n}\sum_{i=1}^{n}\theta_i,\quad S_\theta^2=\frac{1}{n-1}\sum_{i=1}^{n}(\theta_i-\bar{\theta})^2$$

然后令其分别等于Beta分布 $Be(a,b)$ 的数学期望与方差，即

$$\begin{cases}\dfrac{a}{a+b}=\bar{\theta}\\[2ex]\dfrac{ab}{(a+b)^2(a+b+1)}=S_\theta^2\end{cases}$$

解之，即可得超参数 a 与 b 的矩法估计值：

$$\hat{a} = \bar{\theta}\left(\frac{(1-\bar{\theta})\bar{\theta}}{S_{\theta}^{2}} - 1\right), \quad \hat{b} = (1-\bar{\theta})\left(\frac{(1-\bar{\theta})\bar{\theta}}{S_{\theta}^{2}} - 1\right)$$

b）假如根据先验信息只能获得先验均值 $\bar{\theta}$. 可令

$$\frac{a}{a+b} = \bar{\theta}$$

一个方程不能确定两个超参数，譬如 $\bar{\theta}=0.4$，那么满足方程 $a/(a+b)=0.4$ 的 a 与 b 有无穷多组解. 表 5.5 列出若干组，从此表可见，它们的方差 $\mathrm{Var}(\theta)$ 随着 $a+b$ 的增大而减小. 方差减小意味着概率在向均值 $\mathrm{E}(\theta)$ 集中，从而提高 $\mathrm{E}(\theta)=0.4$ 的确信程度. 这样一来，选择 $a+b$ 的问题转化为决策人对 $\mathrm{E}(\theta)=0.4$ 的确信程度大小的问题，若对 $\mathrm{E}(\theta)=0.4$ 很确信，那 $a+b$ 可选得大一些，否则 $a+b$ 就选得小一些. 譬如决策人对 $\mathrm{E}(\theta)=0.4$ 很确信，从而选 $a+b=35$. 从表 5.5 可见，此时 $\hat{a}=14$、$\hat{b}=21$，这样 θ 的先验分布为 $Be(14,21)$.

表 5.5 Beta 分布中超参数与方差的关系

Beta 分布	a	$a+b$	$\mathrm{E}(\theta)$	$\mathrm{Var}(\theta)$
$Be(2,3)$	2	5	0.4	0.0400
$Be(4,6)$	4	10	0.4	0.0218
$Be(8,12)$	8	20	0.4	0.0114
$Be(10,15)$	10	25	0.4	0.0092
$Be(14,21)$	14	35	0.4	0.0067

c）用两个分位数来确定 a 与 b，譬如用两个上下四分位数 θ_u 和 θ_l（见图 5.7）来确定 a 与 b. 从图 5.7 上可以看出，θ_l 与 θ_u 分别满足如下方程

$$\int_0^{\theta_l} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta = 0.25$$

$$\int_{\theta_u}^{1} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta = 0.25$$

这可用 Beta 分布与 F 分布关系和数值积分求解.

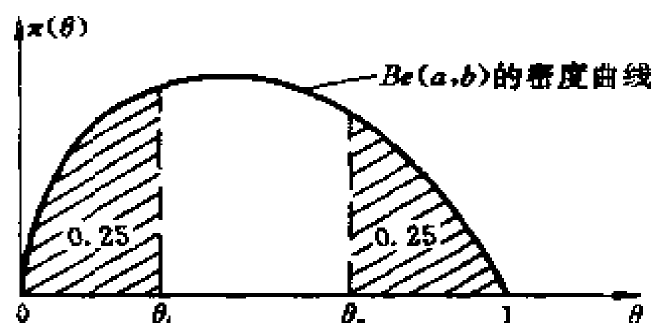


图 5.7　Beta 分布的上,下四分位数

d) 如果决策者对成功概率 θ 一无所知,则可使用 Bayes 假设,即用 Beta 分布 $Be(1,1)$作为 θ 的先验分布. 这种先验分布是所谓无信息先验分布中的一种.

§5.4.5　后验风险准则

设有一个可控参数结构$(\mathscr{X},\mathscr{B},\{p_\theta:\theta\in\Theta\})$,记 $\mathscr{B}_\Theta$ 是参数空间 Θ 中某些子集组成的 σ 代数. 且在$(\Theta,\mathscr{B}_\Theta)$上有一个先验分布 $\pi(\theta)$.

又设$(\Delta,\mathscr{B}_\Delta)$是行动空间. 由$(\mathscr{X},\mathscr{B})$到$(\Delta,\mathscr{B}_\Delta)$上的一切可测变换组成的决策函数类记为 $\mathscr{D}=\{\delta(x)\}$.

最后设有一个定义在 $\Theta\times\Delta$ 上的损失函数 $L(\theta,\delta)$.

这些假设与§5.1.2 中三个基本要素相比只多了一个先验分布. 而这一点正是 Bayes 学派认为在讨论统计问题中是不可缺少的. 有人把他当作统计决策问题的第四个基本要素. 这里把含有先验分布假设的统计决策问题称为 Bayes 决策问题. 前面我们讨论 Bayes 风险准则(§5.4.2)就是在 Bayes 决策问题中进行的. 现在讨论后验风险准则也是在 Bayes 决策问题中进行的.

定义 5.11　在上述假设下,$\mathscr{D}$ 中任一个决策函数 $\delta=\delta(x)$的损失函数 $L(\theta,\delta(x))$对后验分布 $\pi(\theta|x)$的数学期望称为后验风险,记为

$$R_\pi(\delta|x)=\mathrm{E}_{\theta|x}[L(\theta,\delta(x))]$$
$$=\begin{cases}\displaystyle\int_\Theta L(\theta,\delta(x))\pi(\theta|x)\mathrm{d}\theta, & \text{当 }\theta\text{ 为连续量}\\ \displaystyle\sum_i L(\theta_i,\delta(x))\pi(\theta_i|x), & \text{当 }\theta\text{ 为离散量}\end{cases}\tag{5.43}$$

假如在 $\mathscr{D}$ 中存在这样一个决策函数 $\delta^*(x)$，使得

$$R_\pi(\delta^*|x) = \inf_{\delta\in\mathscr{D}} R_\pi(\delta|x), \quad x \in \mathscr{X} \tag{5.44}$$

则称 $\delta^*(x)$为该统计决策问题在后验风险准则下的最优决策函数，或称为 Bayes(后验型)决策函数. 在估计问题中，它又称为 Bayes(后验型)估计.

例 5.29 设 $X=(X_1,\cdots,X_n)$是来自正态总体 $N(\mu,1)$的一个样本. 假如参数 μ 的先验分布取为共轭先验分布 $N(0,\tau^2)$，其中 τ 已知. 而损失函数取为 0-1 损失函数

$$L(\mu,\delta) = \begin{cases} 0, & |\delta-\mu| \leqslant \varepsilon \\ 1, & |\delta-\mu| > \varepsilon \end{cases}$$

要求参数 μ 的 Bayes(后验型)估计.

在例 5.26 中，我们已求得在给定样本观察值 x 下，μ 的后验密度 $\pi(\mu|x)$为正态分布 $N\left(\sum_{i=1}^n x_i/(n+\tau^{-2}),1/(n+\tau^{-2})\right)$. 对任意一个决策函数 $\delta(x)\in\mathscr{D}$，其后验风险为

$$R_\pi(\delta|x) = \int_\Theta L(\mu,\delta)\pi(\mu|x)\mathrm{d}\mu = 1 - P_{\mu|x}(|\delta-\mu| \leqslant \varepsilon|x)$$

要使上述后验风险最小，就要使上式中的条件概率最大. 由于后验分布 $\pi(\mu|x)$是正态分布，要在定长区间(长度为 2ε)上的概率为最大，$\delta(x)$只能取后验分布的均值，即在此场合下，μ 的 Bayes(后验型)估计为

$$\delta_\tau(x) = \frac{\sum_{i=1}^n x_i}{n+\tau^{-2}}$$

它与经典方法得到的估计 $\overline{X}$ 是不同的，一般有

$$\delta_\tau(x) < \overline{X}$$

$\delta_\tau(x)$要比 $\overline{X}$ 更接近于 0，这是由于先验分布给我们带来一些信息之故. 只有当 $\tau\to\infty$时，才有 $\delta_\tau(x)\to\overline{X}$.

例 5.30 设样本 X 只可能来自密度函数 $p_0(x)$或 $p_1(x)$中的一个. 为了研究该样本到底来自哪个分布，我们来考虑如下的简单假设的检验问题：

$$H_0: X \text{ 来自 } p_0(x), \quad H_1: X \text{ 来自 } p_1(x)$$

此时参数空间 $\Theta=\{0,1\}$ 只含两个元素，其中"$\theta=0$"表示 X 来自 $p_0(x)$，"$\theta=1$"表示 X 来自 $p_1(x)$，在 Θ 上的先验分布为

$$P(\theta=0)=\pi_0, P(\theta=1)=\pi_1=1-\pi_0$$

由 Bayes 公式可算得：在给出样本观察值 x 后，θ 的后验分布

$$P(\theta=i|x)=\begin{cases}\dfrac{p_0(x)\pi_0}{p_0(x)\pi_0+p_1(x)\pi_1}, & i=1\\[2ex] \dfrac{p_1(x)\pi_1}{p_0(x)\pi_0+p_1(x)\pi_1}, & i=0\end{cases}$$

在这假设检验问题中也只有接受（用 0 表示）和拒绝（用 1 表示）两个行动，即行动空间 $\Delta=\{0,1\}$. 假如再取损失函数

$$L(i,j)=1-\delta_{i,j}$$

其中 $\delta_{i,j}$ 等于 0 或 1 依赖于 $i\neq j$ 还是 $i=j$. 譬如，$L(0,1)$ 表示当 $\theta=0$ 时，采取拒绝行动引起的损失. 按上式公式 $L(0,1)=1$. 其它场合可类似解释. 在上述假定下，可算得各个行动的后验风险.

$$R_\pi(0|x)=P(\theta=1|x),\quad R_\pi(1|x)=P(\theta=0|x)$$

根据后验风险准则，可以找到如下的 Bayes（后验型）决策函数

$$\delta^{*}(x)=\begin{cases}0, & P(\theta=1|x)<P(\theta=0|x)\\ 1, & P(\theta=1|x)\geqslant P(\theta=0|x)\end{cases}$$

或

$$\delta^{*}(x)=\begin{cases}0, & p_1(x)\pi_1<p_0(x)\pi_0\\ 1, & p_1(x)\pi_1\geqslant p_0(x)\pi_0\end{cases}$$

这个 Bayes（后验型）决策函数所决定的拒绝域（当 $p_0(x)\neq 0$）

$$W=\left\{x:\ \frac{p_1(x)}{p_0(x)}\geqslant\frac{\pi_0}{\pi_1}\right\}$$

与先验分布中两概率的比值有关. 其形式与 Neyman-Pearson 引理给出的拒绝域一样. 这说明在简单假设检验场合下 Bayes（后验型）决策函数就是最优势（MP）检验.

最后我们来研究 Bayes 决策函数 $\delta^{*}(x)$ 与 Bayes（后验型）决策函数 $\delta^{**}(x)$ 的等价性.

定理 5.5　对给定的统计决策问题（包括先验分布的给定）和决

策函数类 $\mathscr{D}$，当 Bayes 风险满足如下条件

$$\inf_{\delta \in \mathscr{D}} R_\pi(\delta) < \infty \tag{5.45}$$

则 Bayes 决策函数 δ^* 与 Bayes（后验型）决策函数 δ^{**} 是等价的. 即使后验风险最小的决策函数 δ^{**} 同时也使 Bayes 风险最小. 反之，使 Bayes 风险最小的决策函数 δ^* 同时也使后验风险最小.

证明：下面仅对 X 和 θ 都是连续随机变量时给出证明. 因为当把 Lebesgue 测度改为各自的控制测度时，不难看出，下述证明仍然成立. 首先利用上述条件(5.45)建立 Bayes 风险与后验风险之间的关系. 对任一个 $\delta \in \mathscr{D}$，根据 Bayes 风险 $R_\pi(\delta)$ 和风险函数 $R(\theta,\delta)$ 的定义，有

$$R_\pi(\delta) = \int_\Theta R(\theta,\delta)\pi(\theta)\mathrm{d}\theta = \int_\Theta \left\{ \int_{\mathscr{X}} L(\theta,\delta)p(x|\theta)\mathrm{d}x \right\}\pi(\theta)\mathrm{d}\theta$$

由于 $p(x|\theta)\pi(\theta)=\pi(\theta|x)m(x)$ 和给定的条件(5.52)，上述两个积分次序可以交换. 于是

$$R_\pi(\delta) = \int_{\mathscr{X}} \left\{ \int_\Theta L(\theta,\delta)\pi(\theta|x)\mathrm{d}\theta \right\} m(x)\mathrm{d}x = \int_{\mathscr{X}} R_\pi(\delta|x)m(x)\mathrm{d}x$$

其中 $R_\pi(\delta|x)$ 是 δ 的后验风险. 这表明：Bayes 风险是后验风险对边际分布 $m(x)$ 的数学期望.

根据 Bayes 决策函数 δ^* 的定义和法都引理，有

$$\begin{aligned} R_\pi(\delta^*) = \inf_{\delta \in \mathscr{D}} R_\pi(\delta) &= \inf_{\delta \in \mathscr{D}} \int_{\mathscr{X}} R_\pi(\delta|x)m(x)\mathrm{d}x \\ &\geqslant \int_{\mathscr{X}} \inf_{\delta \in \mathscr{D}} R_\pi(\delta|x)m(x)\mathrm{d}x \\ &= \int_{\mathscr{X}} R_\pi(\delta^{**}|x)m(x)\mathrm{d}x = R_\pi(\delta^{**}) \\ &\geqslant \inf_{\delta \in \mathscr{D}} R_\pi(\delta) = R_\pi(\delta^*) \end{aligned}$$

由此可得

$$R_\pi(\delta^{**}) = \inf_{\delta \in \mathscr{D}} R_\pi(\delta) = R_\pi(\delta^*)$$

这表明：使后验风险最小的决策函数 δ^{**} 同时也使 Bayes 风险最小.

另外，由上面还可推得

$$\int_{\mathscr{X}} [R_\pi(\delta^*|x) - R_\pi(\delta^{**}|x)]m(x)\mathrm{d}x = 0$$

由 δ^{**} 的定义可知，$R_\pi(\delta^*|x)-R_\pi(\delta^{**}|x)\geqslant 0$. 故得

$$R_\pi(\delta^*|x)=R_\pi(\delta^{**}|x)=\inf_{\delta\in\mathscr{D}}R_\pi(\delta|x)$$

这表明：使 Bayes 风险最小的决策函数 δ^* 也使后验风险最小. 证毕.

应该指出，当定理 5.5 的条件(5.45)满足时，寻求 Bayes 决策函数有两条途径. 实际中，人们常使用后验风险途径. 因为它的计算相对简单和方便. 而当定理 5.5 的条件(5.45)不满足时，即当 $\inf\limits_{\delta\in\mathscr{D}}R_\pi(\delta)=\infty$ 时，两种 Bayes 决策函数就不一定等价. 无论从理论或实践的角度看，使 Bayes 风险为无穷大的决策函数是意义不大的. 我们的注意力应集中在使 Bayes 风险为有限的那些决策函数上. 且人们总希望两种 Bayes 决策函数能够一致. 为此目的，今后只限于考虑条件(5.45)得到满足的情况，这就保证两种 Bayes 决策函数是等价的，不用区分了.

§ 5.5　Bayes 分析

现在我们转向讨论在 Bayes 决策问题(§ 5.4.5)中使用后验期望准则的某些标准应用. 中间还要介绍一些先验分布的确定方法.

§ 5.5.1　Bayes 估计

在 Bayes 决策问题中用来表示未知参数 θ 的点估计的决策函数 $\delta(x)$ 称为 θ 的 Bayes 估计，记为 $\delta^\pi(x)$，其中 π 表示所使用的先验分布. 在常用损失函数下，Bayes 估计有如下几个结论.

定理 5.6　在给定先验分布 $\pi(\theta)$ 和平方损失 $L(\theta,\delta)=(\delta-\theta)^2$ 下，θ 的 Bayes 估计 $\delta^\pi(x)$ 为后验分布 $\pi(\theta|x)$ 的均值，即 $\delta^\pi(x)=\mathrm{E}(\theta|x)$.

证明：在平方损失下，任一个决策函数 $\delta=\delta(x)$ 的后验风险为

$$\mathrm{E}[(\delta-\theta)^2|x]=\delta^2-2\delta\mathrm{E}(\theta|x)+\mathrm{E}(\theta^2|x)$$

此后验风险达到最小，仅在 $\delta^\pi(x)=\mathrm{E}(\theta|x)$，证毕.

定理 5.7　在给定先验分布 $\pi(\theta)$ 和加权平方损失函数 $L(\theta,\delta)=\lambda(\theta)(\delta-\theta)^2$ 下，θ 的 Bayes 估计为

$$\delta^\pi(x)=\frac{\mathrm{E}[\lambda(\theta)\theta|x]}{\mathrm{E}[\lambda(\theta)|x]}$$

证明：类似于定理 5.6 的证明即可得，证毕.

定理 5.8 在向量参数 $\boldsymbol{\theta}=(\theta_1,\cdots,\theta_p)'$ 的场合下，对给定的先验分布 $\pi(\theta)$ 和二次损失函数 $L(\boldsymbol{\theta},\boldsymbol{\delta})=(\boldsymbol{\delta}-\boldsymbol{\theta})'\boldsymbol{Q}(\delta-\theta)$，$\boldsymbol{Q}$ 为正定阵，$\boldsymbol{\theta}$ 的 Bayes 估计为后验分布 $\pi(\theta|x)$ 的均值向量，即

$$\delta^\pi(x)=\mathrm{E}[\theta|x]=\begin{pmatrix}\mathrm{E}(\theta_1|x)\\ \vdots\\ \mathrm{E}(\theta_p|x)\end{pmatrix}$$

这个结论表明：在正定二次损失下，θ 的 Bayes 估计不受正定阵 $\boldsymbol{Q}$ 的选取的干扰，这一特性常被称为 $\boldsymbol{\theta}$ 的 Bayes 估计 δ^π 关于 $\boldsymbol{Q}$ 是稳健的.

证明：在二次损失下，任一个决策函数向量 $\delta(x)=(\delta_1(x),\cdots,\delta_p(x))'$ 的后验风险为

$$\begin{aligned}&\mathrm{E}[(\boldsymbol{\delta}-\boldsymbol{\theta})'\boldsymbol{Q}(\boldsymbol{\delta}-\boldsymbol{\theta})|x]\\ &=\mathrm{E}[((\boldsymbol{\delta}-\boldsymbol{\delta}^\pi)+(\boldsymbol{\delta}^\pi-\boldsymbol{\theta}))'\boldsymbol{Q}((\boldsymbol{\delta}-\boldsymbol{\delta}^\pi)+(\boldsymbol{\delta}^\pi-\boldsymbol{\theta}))|x]\\ &=(\boldsymbol{\delta}-\boldsymbol{\delta}^\pi)'\boldsymbol{Q}(\boldsymbol{\delta}-\boldsymbol{\delta}^\pi)'+\mathrm{E}[(\boldsymbol{\delta}^\pi-\boldsymbol{\theta})'\boldsymbol{Q}(\boldsymbol{\delta}^\pi-\boldsymbol{\theta})|x]\end{aligned}$$

上述最后一个等式是考虑到 $\mathrm{E}(\boldsymbol{\delta}^\pi-\boldsymbol{\theta}|x)=0$. 上式第二项为常量，而第一项非负，故使上式最小仅在 $\boldsymbol{\delta}=\boldsymbol{\delta}^\pi(x)$ 达到. 证毕.

例 5.31 设 $X_1,\cdots,X_n$ 是来自正态总体 $N(\theta,\sigma^2)$ 的一个样本，其中 σ^2 已知，若 θ 的先验分布取其共轭先验分布 $N(\mu,\tau)$，其中 μ,τ 已知，则在例 5.26 中算得其后验分布为 $N(\mu_1,\sigma_1^2)$，其中

$$\mu_1=\frac{\tau^{-2}\mu+\sigma_0^{-2}\bar{x}}{\tau^{-2}+\sigma_0^{-2}},\quad \sigma_1^2=(\tau^{-2}+\sigma_0^{-2})^{-1},\quad \sigma_0^2=\frac{\sigma^2}{n}$$

若取平方损失函数，则 θ 的 Bayes 估计为其后验均值，即

$$\delta^\pi(x)=\lambda\mu+(1-\lambda)\bar{x}$$

其中 $\lambda=\tau^{-2}/(\tau^{-2}+\sigma_0^{-2})$，当 $\sigma_0^2>\tau^2$ 时，$\lambda>\frac{1}{2}$，于是从上式知，先验均值 μ 在 Bayes 估计中占的比重大一些，当 $\sigma_0^2<\tau^2$ 时，$\lambda<\frac{1}{2}$，那么样本均值 $\bar{x}$ 在 Bayes 估计中占的比重大一些，这是符合人们的直观认识："方差小的信息更应受到重视".

作为数值例子，我们对一个儿童做智力测验. 设测验结果 $X\sim N(\theta,100)$，其中 θ 为该儿童智商的真值. 又设 θ 的先验分布为 $N(100,225)$. 应用上述结果，在 $n=1$ 时，可在 $X=x$ 条件下获得 θ 的后验分布 $N(\mu_1,\sigma_1^2)$，其中

$$\mu_1=\frac{400+9x}{13},\quad \sigma_1^2=(8.32)^2$$

假如这个儿童智商测验得分为 115 分，则他的智商的 Bayes 估计为 $\delta^{\pi}(115)=(400+9\times 115)/13=110.38$.

例 5.32 经过早期筛选(出厂前的检验)后彩色电视接收机(简称彩电)的寿命服从指数分布，其密度函数为

$$p(t|\theta)=\theta^{-1}\mathrm{e}^{-t/\theta},\quad t>0$$

其中 $\theta>0$ 是彩电的平均寿命. 若从一批彩电中随机抽取 n 台进行寿命试验，试验到第 r 台失效为止，获得截尾样本：$t_1\leqslant t_2\leqslant\cdots\leqslant t_r$. 此截尾样本 $\boldsymbol{t}=(t_1,\cdots,t_r)$ 的联合密度函数为

$$p(t|\theta)\propto\prod\nolimits_{i=1}^{r}(\theta^{-1}\mathrm{e}^{-t_i/\theta})[\mathrm{e}^{-t_r/\theta}]^{n-r}=\theta^{-r}\exp\{-T_r/\theta\}$$

其中 $T_r=t_1+\cdots+t_r+(n-r)t_r$ 称为总试验时间，为寻求彩电平均寿命 θ 的 Bayes 估计，要确定先验分布. 据国内外的经验，选用倒 Gamma 分布 $IG(\alpha,\lambda)$ 作为 θ 的先验分布是恰当的. $IG(\alpha,\lambda)$ 的密度函数为

$$\pi(\theta)=\frac{\lambda^{\alpha}}{\Gamma(\alpha)}\theta^{-(\alpha+1)}\mathrm{e}^{-\lambda/\theta},\quad \theta>0$$

其中 $\alpha>0,\lambda>0$ 是两个待定参数. 它的数学期望 $\mathrm{E}(\theta)=\lambda/(\alpha-1)$.

把先验分布 $\pi(\theta)$ 与截尾样本分布相乘可得后验分布的核

$$\pi(\theta|t)\propto p(t|\theta)\pi(\theta)\propto\theta^{-(\alpha+r+1)}\mathrm{e}^{-(\lambda+T_r)/\theta}$$

这仍然是倒 Gamma 分布的核，故后验分布 $\pi(\theta|t)$ 是 $IG(\alpha+r,\lambda+T_r)$. 假如取平方损失函数，则 θ 的 Bayes 估计为后验均值

$$\hat{\theta}=\frac{\lambda+T_r}{\alpha+r-1}$$

为了最后确定这个估计，有关单位收集先验信息和抽样信息. 我国彩电生产厂做了大量的彩电寿命试验，仅 15 个工厂试验室和一些独立试验室就对 13 142 台彩电进行了共计 5 369 812 台时试验，而且还对 9240

台彩电进行了三年现场跟踪试验，总共进行了 5 547 810 台时试验. 这两类试验总共失效台数不超过 250 台. 对如此大量先验信息加工整理后，确认我国彩电平均寿命不低于 30 000 小时；它的 10%分位数大约为 11 250 小时，经过一些专家认定，这两个数据是符合我国八十年代中期彩电寿命的实际情况，也是留有余地的. 由此可列出如下两个方程

$$\begin{cases} \dfrac{\lambda}{\alpha - 1} = 30\,000 \\ \displaystyle\int_0^{11\,250} \pi(\theta)\mathrm{d}\theta = 0.1 \end{cases}$$

在计算机上解此方程组，得

$$\alpha = 1.956, \quad \lambda = 2868$$

这样一来，我们就完全确定了先验分布 $IG(1.956, 2868)$，假如随机抽取 100 台彩电进行 400 小时试验，没有一台失效. 这时总试验时间 $T_r = 100 \times 400 = 40\,000$ 小时，$r=0$，于是彩电平均寿命 θ 的 Bayes 估计为

$$\hat{\theta} = \frac{2.868 + T_r}{1.956 + r - 1} = \frac{42\,868}{0.956} = 44\,841(\mathrm{h})$$

定理 5.9　在给定先验分布 $\pi(\theta)$ 和绝对损失函数 $L(\theta,\delta) = |\delta - \theta|$ 下，θ 的 Bayes 估计为 $\delta^\pi(x)$ 后验分布 $\pi(\theta|x)$ 的中位数.

证明：设 m 为 $\pi(\theta|x)$ 的中位数，又设 $\delta = \delta(x)$ 为 θ 的另一估计. 为确定起见，先设 $\delta > m$. 由绝对损失函数定义可得

$$L(\theta,m) - L(\theta,\delta) = \begin{cases} m - \delta, & \theta \leqslant m \\ 2\theta - (m + \delta), & m < \theta < \delta \\ \delta - m, & \theta \geqslant \delta \end{cases}$$

当 $m < \theta < \delta$ 时，上式中 $2\theta - (m + \delta) \leqslant 2\delta - (m + \delta) = \delta - m$. 所以上式为

$$L(\theta,m) - L(\theta,\delta) \leqslant \begin{cases} m - \delta, & \theta \leqslant m \\ \delta - m, & \theta > m \end{cases}$$

由中位数定义知 $P(\theta \leqslant m | x) \geqslant 1/2$ 而 $P(\theta > m | x) < 1/2$. 由此可得

$$\begin{aligned} & R_\pi(m|x) - R_\pi(\delta|x) \\ & = E_{\theta|x}[L(\theta,m) - L(\theta,\delta)] \end{aligned}$$

$$\leqslant (m-\delta)P(\theta \leqslant m|x) + (\delta - m)P(\theta > m|x)$$
$$\leqslant (m-\delta)/2 + (\delta - m)/2 = 0$$

于是对 $\delta > m$ 有

$$R_\pi(m|x) \leqslant R_\pi(\delta|x)$$

类似地，对 $\delta < m$ 亦可证得上述不等式成立. 这就表明后验分布中位数 m 是使后验风险最小，故 m 是其 Bayes 估计，证毕.

定理 5.10 在给定先验分布 $\pi(\theta)$ 和 0-1 损失函数

$$L(\theta,\delta) = \begin{cases} 0, & |\delta - \theta| \leqslant \varepsilon \\ 1, & |\delta - \theta| > \varepsilon \end{cases}$$

当 ε 较小时，θ 的 Bayes 估计为后验分布 $\pi(\theta|x)$ 的众数.

证明：下面对 θ 为连续随机变量场合给出证明. 在 θ 为离散随机变量亦可类似证明. 考察 θ 的任一估计 $\delta(x)$ 的后验风险

$$R_\pi(\delta|x) = \int_\Theta L(\theta,\delta)\pi(\theta|x)\mathrm{d}\theta = 1 - \int_{\delta-\varepsilon}^{\delta+\varepsilon}\pi(\theta|x)\mathrm{d}\theta$$

要使上述后验风险最小，就要使上式右端积分最大，这只要在 ε 很小时，δ 取被积函数 $\pi(\theta|x)$ 的极大值就可达到. 故取 $\pi(\theta|x)$ 的众数作为 θ 的估计就可使后验风险最小，证毕.

例 5.33 设随机变量 X 服从几何分布

$$P(X = x) = \theta(1-\theta)^x, \quad x = 0,1,\cdots$$

其中参数 θ 只能取 1/4，2/4 和 3/4 三个值，并以相同概率取这三个值，如今只获得一个观察值 $X=2$，要在 0-1 损失函数下寻求 θ 的 Bayes 估计.

在这个问题中，$X=2$ 的概率为 $P(X=2|\theta)=\theta(1-\theta)^2$，而 θ 为 1/4，2/4，3/4的概率皆为 1/3. 所以联合概率

$$P\left(X = 2,\theta = \frac{i}{4}\right) = \frac{1}{3}\cdot\frac{i}{4}\left(1-\frac{i}{4}\right)^2, \quad i = 1,2,3$$

可以算得无条件概率 $P(X=2)=5/48$. 于是在 $X=2$ 条件下，θ 的后验分布为

$$P\left(\theta = \frac{i}{4}\Big|X = 2\right) = \frac{4i}{5}\left(1-\frac{i}{4}\right)^2, \quad i = 1,2,3$$

可以算得：$\theta=1/4$ 的后验概率为 9/20，θ 为另外两个值的后验概率较

小，分别为 8/20 和 3/20. 所以在 0-1 损失函数下，θ 的 Bayes 估计为 1/4.

评价 Bayes 估计 $\delta^\pi(x)$ 的精度常用后验均方差

$$\mathrm{MSE}(\delta^\pi|x) = \mathrm{E}_{\theta|x}(\delta^\pi - \theta)^2$$

表示，或用其平方根 $(\mathrm{MSE}(\delta^\pi|x))^{1/2}$（称为标准误）表示. 容易算得

$$\mathrm{MSE}(\delta^\pi|x) = \mathrm{Var}(\delta^\pi|x) + [\delta^\pi(x) - \mathrm{E}(\theta|x)]^2$$

可见，当 Bayes 估计 $\delta^\pi(x)$ 为后验均值时，Bayes 估计的精度就用 δ^π 的后验方差 $\mathrm{Var}(\delta^\pi|x)$ 表示，或用后验标准差 $(\mathrm{Var}(\delta^\pi|x))^{1/2}$ 表示.

譬如在讨论儿童智商测验的例 5.31 中，在一次测验值 x 下获得后验分布 $\pi(\theta|x)=N((400+9x)/13, 900/13)$. 当 $x=115$ 时，其后验分布为 $N(110.39, (8.32)^2)$. 若取后验均值为该儿童智商 θ 的 Bayes 估计，即 $\delta^\pi=110.39$，则其后验标准差为 8.32. 若用经典估计 $\hat{\theta}=x=115$ 时，其后验均方差为

$$\mathrm{MSE}(\delta^\pi|x) = (8.23)^2 + (110.39 - 115)^2 = 90.48$$

其标准误为 $\sqrt{90.48}=9.49$.

§5.5.2 Bayes 估计的性质

我们先讨论 Bayes 估计的容许性.

定理 5.11 在 Bayes 决策问题中，假如先验分布 $\pi(\theta)$ 在 Θ 上处处为正. 对每一个估计 $\delta(x)$，其风险函数 $R(\theta,\delta)$ 是 θ 的连续函数，其 Bayes 风险 $R_\pi(\delta)$ 是有限的，则 θ 的 Bayes 估计 δ^π 是容许的.

证明： 倘若 δ^π 是非容许的，则存在另一个估计 $\delta(x)$，使得

$$R(\theta,\delta) \leqslant R(\theta,\delta^\pi), \quad \forall\, \theta \in \Theta$$

且至少对某个 $\theta_1 \in \Theta$，使得上述严格不等式成立，即

$$R(\theta_1,\delta) < R(\theta_1,\delta^\pi)$$

根据连续性的假设，存在一个正数 ε 及邻域 S_ε，使得

$$R(\theta,\delta) < R(\theta,\delta^\pi) - \varepsilon, \quad \forall\, \theta \in S_\varepsilon$$

上式两边对先验分布函数 $H(\theta)$ 求积分，可得

$$R_\pi(\delta) = \int_{S_\varepsilon} R(\theta,\delta)H(\mathrm{d}\theta) + \int_{\overline{S}_\varepsilon} R(\theta,\delta)H(\mathrm{d}\theta)$$

$$\leqslant \int_{S_\epsilon}[R(\theta,\delta)-\epsilon]H(\mathrm{d}\theta)+\int_{\bar{S}_\epsilon}R(\theta,\delta)H(\mathrm{d}\theta)$$

$$=R_\pi(\delta^\pi)-\epsilon H(S_\epsilon)<R_\pi(\delta^\pi)$$

最后一个不等式成立是由于 $H(S_\epsilon)>0$. 这与 δ^π 是 Bayes 估计矛盾. 所以 δ^π 是容许的. 证毕.

这个定理虽然简单,却有很大实用价值. 首先它可说明一大批 Bayes 估计是容许的. 其次它指出证明"一个估计是容许估计"的一条思路. 假如对给定的估计 $\delta(x)$,能找到满足定理 5.11 的先验分布 $\pi(\theta)$,使得 $\delta(x)$ 就是关于此先验分布下的 Bayes 估计即可.

定理 5.12　在给定的 Bayes 决策问题中,假如对给定的先验分布 $\pi(\theta)$,θ 的 Bayes 估计 $\delta^\pi(x)$ 是唯一的,则它也是容许的.

证明: 倘若 $\delta^\pi(x)$ 是非容许的,则存在另一个估计 $\delta(x)\neq\delta^\pi(x)$,使得

$$R(\theta,\delta)\leqslant R(\theta,\delta^\pi),\quad \forall\, \theta\in\Theta$$

且至少对某一个 θ 有严格不等式成立. 上式两边对先验分布积分,立刻可看出,$\delta(x)$ 亦应是 θ 的 Bayes 估计. 这与唯一性矛盾. 故 $\delta^\pi(x)$ 是容许的,证毕.

从这个定理可以看出,当损失函数是严格凸函数时,其 Bayes 估计必是唯一的,从而也是容许的.

下面我们转入讨论 Bayes 估计与最小最大估计的关系.

定理 5.13　在给定的 Bayes 决策问题中,$\mathscr{D}$ 是非随机决策函数类. 假如 $\delta^*(x)\in\mathscr{D}$ 是对应先验分布 $H^*(\theta)$ 的 Bayes 估计. 且其风险函数 $R(\theta,\delta^*)=\rho^*$(常数),　$\forall\, \theta\in\Theta$,则 δ^* 是参数 θ 的最小最大估计.

证明: 因 δ^* 是对应先验分布 $H^*(\theta)$ 的 Bayes 估计,故 δ^* 可使后验风险和 Bayes 风险达到最小. 若记 $\mathscr{H}$ 为 Θ 上一切先验分布组成的分布族,则有

$$\rho^*=\int R(\theta,\delta^*)H^*(\mathrm{d}\theta)=\inf_{\delta\in\mathscr{D}}\int R(\theta,\delta)H^*(\mathrm{d}\theta)$$

$$\leqslant \sup_{H\in\mathscr{H}}\inf_{\delta\in\mathscr{D}}\int R(\theta,\delta)H(\mathrm{d}\theta)=\sup_{H\in\mathscr{H}}\inf_{\delta\in\mathscr{D}}R_H(\delta)$$

$$\leqslant \inf_{\delta \in \mathscr{D}} \sup_{H \in \mathscr{H}} R_H(\delta) \leqslant \inf_{\delta \in \mathscr{D}} \sup_{\theta \in \Theta} R(\theta, \delta)$$

上式倒数第二个不等式成立是因为"从最小中取最大的总不会超过从最大中取最小的"之故. 事实上,对每个 $\delta \in \mathscr{D}$,总有

$$R_H(\delta) \leqslant \sup_{H \in \mathscr{H}} R_H(\delta)$$

$$\inf_{\delta \in \mathscr{D}} R_H(\delta) \leqslant \inf_{\delta \in \mathscr{D}} \sup_{H \in \mathscr{H}} R_H(\delta)$$

$$\sup_{H \in \mathscr{H}} \inf_{\delta \in \mathscr{D}} R_H(\delta) \leqslant \inf_{\delta \in \mathscr{D}} \sup_{H \in \mathscr{H}} R_H(\delta)$$

上式最后一个不等式成立是由于对任一个 $H \in \mathscr{H}$,总有

$$R_H(\delta) = \int_\Theta R(\theta, \delta) H(\mathrm{d}\theta) \leqslant \sup_{\theta \in \Theta} R(\theta, \delta)$$

$$\sup_{H \in \mathscr{H}} R_H(\delta) \leqslant \sup_{\theta \in \Theta} R(\theta, \delta)$$

$$\inf_{\delta \in \mathscr{D}} \sup_{H \in \mathscr{H}} R_H(\delta) \leqslant \inf_{\delta \in \mathscr{D}} \sup_{\theta \in \Theta} R(\theta, \delta)$$

另一方面,由于 $R(\theta, \delta^*)$ 在 Θ 上为常数 ρ^*,故有

$$\rho^* = \sup_{\theta \in \Theta} R(\theta, \delta^*) \geqslant \inf_{\delta \in \mathscr{D}} \sup_{\theta \in \Theta} R(\theta, \delta)$$

比较上面两个方向不同的不等式,即证得 δ^* 是 θ 的最小最大估计. 证毕.

例 5.34 设随机变量 $X \sim b(n, \theta)$, $0 < \theta < 1$. 若 θ 的先验分布取 Beta 分布 $Be\left(\sqrt{n}/2, \sqrt{n}/2\right)$. 若从二项分布仅获得一个观察值 x,则其后验分布为 Beta 分布 $Be\left(x + \sqrt{n}/2, n - x + \sqrt{n}/2\right)$. 于是在平方损失函数下,$\theta$ 的 Bayes 估计为

$$\delta^\pi(x) = \frac{x + \sqrt{n}/2}{n + \sqrt{n}}$$

由于先验分布 $Be\left(\sqrt{n}/2, \sqrt{n}/2\right)$ 在 $(0,1)$ 上处处为正. 由定理 5.11 知,此 Bayes 估计 $\delta^\pi(x)$ 是容许的,又其风险函数为

$$R(\theta, \delta^\pi) = \mathrm{E}_{x|\theta}\left[\frac{x + \sqrt{n}/2}{n + \sqrt{n}} - \theta\right]^2 = \frac{1}{4(1 + \sqrt{n})^2}$$

它与 θ 无关,由定理 5.13 知,δ^π 还是最小最大估计.

有趣的是把 δ^π 与 θ 的 MLE $\hat{\theta} = x/n$ 进行比较. 容易算得 $\hat{\theta}$ 的风险函数

$$R(\theta,\hat{\theta}) = \mathrm{E}_{x|\theta}\left[\frac{x}{n} - \theta\right]^2 = \frac{\theta(1-\theta)}{n}$$

图 5.8 上画出它们的风险函数图形. 从图上可见, 在均方误差标准下, 极大似然估计 $\hat{\theta}$(也是最小方差无偏估计)仅在两端(θ 靠近 0 或靠近 1 时)优于 δ^π, 而在中段(0.5 附近)不如 δ^π. 特别在 n 不太大时, 中段范围 $2\theta_n$ 还不小, 譬如 $n=25$ 时, $2\theta_n=0.55$, $n=100$ 时, $2\theta_n=0.42$.

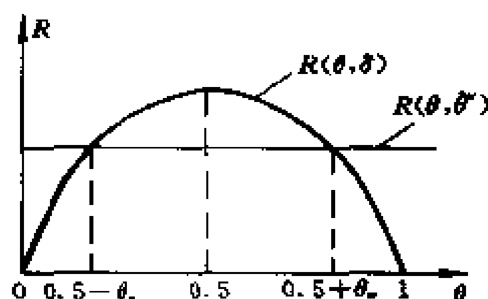


图 5.8 两个风险函数比较.

$$\theta_n = \frac{\sqrt{1+2\sqrt{2}}}{2(1+\sqrt{n})}$$

定理 5.14 在给定的 Bayes 决策问题中, 设 $\{H_k : k \geqslant 1\}$ 是 Θ 上的一个先验分布列. $\{\delta_k : k \geqslant 1\}$ 和 $\{R_{H_k}(\delta_k) : k \geqslant 1\}$ 是对应的 Bayes 估计列和 Bayes 风险列. 假如 δ^* 是 θ 的一个估计, 且它的风险函数满足

$$\sup_{\theta\in\Theta} R(\theta,\delta^*) \leqslant \lim_{k\to\infty} \sup R_{H_k}(\delta_k)$$

则 δ^* 是 θ 的最小最大估计.

证明: 倘若 δ^* 不是 θ 的最小最大估计, 则存在这样一个估计 $\tilde{\delta}$, 使得

$$\sup_{\theta\in\Theta} R(\theta,\tilde{\delta}) < \sup_{\theta\in\Theta} R(\theta,\delta^*)$$

由于 δ_k 是对应先验分布 $H_k(\theta)$ 的 Bayes 估计($k\geqslant 1$), 故其 Bayes 风险最小, 从而

$$R_{H_k}(\delta_k) \leqslant R_{H_k}(\tilde{\delta}) = \int_\Theta R(\theta,\tilde{\delta})H_k(\mathrm{d}\theta) \leqslant \sup_{\theta\in\Theta} R(\theta,\tilde{\delta})$$

还有

$$R_{H_k}(\delta_k) \leqslant \sup_{\theta\in\Theta} R(\theta,\tilde{\delta}) < \sup_{\theta\in\Theta} R(\theta,\delta^*)$$

这与假设的条件矛盾, 这说明 δ^* 是 θ 的最小最大估计, 证毕.

定理 5.15 在给定的 Bayes 决策问题中, 若 θ 的一个估计 δ^* 的风险函数 $R(\theta,\delta^*)$ 在 Θ 上为常数 ρ^*, 且存在一个先验分布列 $\{H_k\}$ 使得相应 Bayes 估计 δ_k 的 Bayes 风险满足

$$\lim_{k\to\infty} R_{H_k}(\delta_k) = \rho^*$$

则 δ^* 是 θ 的最小最大估计.

证明：对任意的 $\theta\in\Theta$，有

$$\sup_{\theta\in\Theta}R(\theta,\delta^*)=R(\theta,\delta^*)=\rho^*=\lim_{k\to\infty}R_{H_k}(\delta_k).$$

这表明定理 5.14 的条件满足，故 δ^* 为 θ 的最小最大估计，证毕.

例 5.35 设 $X_1,\cdots,X_n$ 是来自正态总体 $N(\mu,1)$ 的一个样本. 则在 0-1 损失函数下利用定理 5.15 可以证明：样本均值 $\overline{X}_n$ 是 μ 的最小最大估计.

应用定理 5.15 的关键在于选取适当的先验分布列，在例 5.27 中，曾选用 $N(0,\tau^2)$ 作为 μ 的先验分布，并在 0-1 损失下获得 μ 的 Bayes 估计

$$\delta_\tau(x)=\frac{\sum X_i}{n+\tau^{-2}}=\frac{\overline{X}_n}{1+1/n\tau^2}$$

容易看出，δ_τ 仍服从正态分布，其期望与方差分别为

$$\mathrm{E}(\delta_\tau)=\frac{\mu}{1+1/n\tau^2},\quad \mathrm{Var}(\delta_\tau)=n^{-1}\left(1+\frac{1}{n\tau^2}\right)^{-2}$$

它的风险函数

$$\begin{aligned}R(\mu,\delta_\tau)&=P(|\delta_\tau-\mu|\geqslant\varepsilon)=1-P(\mu-\varepsilon<\delta_\tau<\mu+\varepsilon)\\&=2-\Phi\left(\sqrt{n}\left[\varepsilon\left(1+\frac{1}{n\tau^2}\right)+\frac{\mu}{n\tau^2}\right]\right)-\\&\quad\Phi\left(\sqrt{n}\left[\varepsilon\left(1+\frac{1}{n\tau^2}\right)-\frac{\mu}{n\tau^2}\right]\right)\end{aligned}$$

在此基础上，我们选一个序列 $\tau_1<\tau_2<\cdots\to\infty$，从而获得一先验分布列 $\{N(0,\tau_i^2)\}$，一 Bayes 估计列 $\{\delta_{\tau_i}\}$ 和一风险函数列 $\{R(\mu,\delta_{\tau_i})\}$. 由诸 $R(\mu,\delta_{\tau_i})<2$，根据 Lebesgue 控制收敛定理有

$$\begin{aligned}\lim_{i\to\infty}R_{H_{\tau_i}}(\delta_{\tau_i})&=\lim_{i\to\infty}\mathrm{E}_{H_{\tau_i}}\left[R(\mu,\delta_{\tau_i})\right]\\&=\mathrm{E}_{H_{\tau_i}}\mathrm{E}\left[\lim_{i\to\infty}R(\mu,\delta_{\tau_i})\right]\\&=\mathrm{E}_{H_{\tau_i}}\left[2-2\Phi(\sqrt{n}\,\varepsilon)\right]\\&=2\left[1-\Phi(\sqrt{n}\,\varepsilon)\right]=\rho^*\end{aligned}$$

而 ρ^* 也正是样本均值 $\overline{X}_n$ 在 0-1 损失函数下的风险. 最后由定理 5.15，

知样本均值 $\bar{X}_n$ 在 0-1 损失函数下是 μ 的最小最大估计.

最后在这里简单地提及一个问题:在 Bayes 估计中不用无偏性来评价一个估计的好坏,这是因为在无偏估计的定义中,$E\hat{\theta}(x)=\theta$,其中数学期望是对样本空间中所有可能样本 x 而求的,可实际中绝大多数样本尚未出现过,甚至重复数百次也不会出现的样本也要在评价估计量中占一席之地,这是不应该的. 另一方面,在实际使用中不少估计量只使用一次或数次,所以 Bayes 学派认为,评价一个估计量的好坏只能依据在试验中所收集到观察值,不应该使用尚未观察到的数据. 这一观点被 Bayes 学派称为"条件观点". 据此,估计的无偏性在 Bayes 估计中不予考虑.

Pratt 在 1962 年举了一个例子说明经典学派的"频率观点"在处理类似问题中的看法. 一位工程师在对电子管的一个随机样本测量板极电压时. 所用测量仪器极其精密,以致误差可忽略不记. 一位统计学家检查了测量值,看上去为正态分布,变化范函为 75 到 99 伏特,均值为 87,标准差为 4. 他进行了统计分析,给出了总体均值的置信区间. 后来他到试验室发现,电压计读数至多为 100 伏特. 于是认为总体是"截尾的",应重新处理数据. 但工程师说,他有另一台电压计,同样精度,能测到 1000 伏特. 如果电压超过 100 伏特,就会用这一台测量. 这使频率派统计学家感到放心,因为这表明总体毕竟是完整的. 无须重新处理数据. 第二天工程师打电话说:"我刚刚发现,那台高量程的电压计坏了."统计学家查明,那台高量程的电压计修复之前,试验没有停止. 故通知工程师,数据需要重新分析. 工程师大吃一惊地说:"即使那台高量程的电压计是好的,试验所测数据也会一样. 无论如何,我们得到的是我的样本的精确电压值. 下一次你就要问到我的示波器了吧."

这个例子所讨论的问题涉及两个不同总体. 即是否需要用在 100 处的截尾正态总体,这对均值的置信区间会有很大影响,可见 Bayes 观点认为,没有发生的 x 值(如 x 大于或等于 100)对均值的推断和决策是没有意义的. 从这个例子可见条件观点与频率观点间的差别,也可看出工程师是容易接受条件观点的.

§5.5.3 无信息先验分布

从 Bayes 分析诞生之日开始就伴随一个问题：没有先验信息场合如何确定先验分布？如何进行 Bayes 分析？Bayes 学派对此作了大量研究，获得一批成果，统称为无信息先验分布. 这是 Bayes 学派在最近三十多年来获得的重要成果之一，下面分几点说明.

1. Bayes 假设

人们常把"没有 θ 的任何信息"理解为对 θ 的任何可能值既无偏爱，又同等无知. 因此很自然地把 θ 取值范围上的均匀分布取作 θ 的先验分布. 譬如，若对男婴的出生率，射击的命中率，商品的市场占有率等一无所知时可用(0,1)上的均匀分布 $U(0,1)$ 作为先验分布. 若 $\Theta=(a,b)$，那可用 $U(a,b)$ 作为先验分布. 这种称为"Bayes 假设"的均匀分布是恰当地表达了人们对 θ 的一种认识.

使用 Bayes 假设也会遇到一些麻烦，主要是以下两个.

一是当 Θ 为无限区间时，无法在其上定义一个正常的先验分布，下面的例子具体说明遇到什么样的麻烦.

例 5.36 设总体 $X\sim N(\theta,1)$，其中 $\theta\in(-\infty,\infty)=\Theta$，若对 θ 既无任何信息，也无偏爱，那应取如下"均匀分布"

$$\pi(\theta)=c,\qquad -\infty<\theta<\infty$$

显然，这不是一个正常的概率密度函数，因为 $\int_{-\infty}^{\infty}\pi(\theta)\mathrm{d}\theta=\infty$，值得注意的是，若按 Bayes 公式计算，可得

$$\pi(\theta|x)=\frac{p(x|\theta)\pi(\theta)}{\int_{-\infty}^{\infty}p(x|\theta)\pi(\theta)\mathrm{d}\theta}=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}(\theta-x)^2\right\}$$

即在给定 x 下，θ 的后验分布为 $N(x,1)$，这是一个正常的概率分布，且 c 的选择并不重要，常选用 $\pi(\theta)=1$. 若用后验均值作为 θ 的估计，则有 $\hat{\theta}=x$，这与只用样本信息获得的结果是一致. 这种现象不是个别的. Bayes 学派认为这种不正常的均匀分布也是先验分布，并称之为广义先验分布. 它的一般定义如下.

定义 5.12 设总体 $X\sim p(x|\theta)$，$\theta\in\Theta$，若满足下列条件

(i) $\pi(\theta)>0$ 且 $\int_{\Theta}\pi(\theta)\mathrm{d}\theta=\infty$；

(ii) 由此决定的后验密度 $\pi(\theta|x)$是正常的密度函数.

则称 $\pi(\theta)$为 θ 的广义先验密度.

在例 5.36 中给出的 $\pi(\theta)=1$ 就是正态均值 θ 的一个广义先验分布，也是一个典型的无信息先验分布，并称它为实数集上的均匀分布.

另一个麻烦是 Bayes 假设不满足变换下的不变性. 譬如正态总体 $N(0,\sigma^2)$中的方差 σ^2 与标准差 σ 都在$(0,\infty)$上取值. 若 σ 的先验分布取为 $\pi(\sigma)$，则按概率运算法则，$\eta=\sigma^2$ 的分布应为 $\pi(\sqrt{\eta})/(2\sqrt{\eta})$. 假如对它们均无任何先验信息可用，若 σ 的无信息先验分布被选为常数，那么 $\eta=\sigma^2$ 的无信息先验密度应与 $\eta^{-1/2}$成比例. 这与 Bayes 假设矛盾. 这个矛盾说明，不能随意设定一个常数为某参数的无信息先验分布. 一些研究表明，无信息先验分布的选取与参数在总体分布中的地位很有关系. 以下叙述其中几个常用的方法.

2. 位置参数的无信息先验分布

设总体 X 的密度函数具有形式 $p(x-\theta)$，其样本空间 $\mathscr{X}$ 和参数空间 Θ 皆为实数集 $\mathbf{R}$. 这类密度组成位置参数族，θ 称为位置参数.

设想让 X 移动一个量 c 得到 $Y=X+c$，同时让参数 θ 也移动同一个量 c 得到 $\eta=\theta+c$. 显然 Y 有密度 $p(y-\eta)$，它仍然是位置参数族的成员. 因此 θ 与 η 应具有相同的无信息先验分布，即

$$\pi(\tau)=\pi^*(\tau)$$

其中 $\pi(\cdot)$和 $\pi^*(\cdot)$分别为 θ 和 η 的无信息先验分布. 另一方面，由变换 $\eta=\theta+c$ 可算得 η 的无信息先验密度为

$$\pi^*(\eta)=\left|\frac{\mathrm{d}\theta}{\mathrm{d}\eta}\right|\pi(\eta-c)=\pi(\eta-c)$$

比较上面两式可得 $\pi(\eta)=\pi(\eta-c)$. 此式要对任意 η 的取值都成立. 特别取 $\eta=c$，则有 $\pi(c)=\pi(0)=$常数. 又由于 c 的任意性，故 θ 的无信息先验分布应为 $\pi(\theta)=1$. 这表明：位置参数在位移变换保持不变的无信息先验分布是 $\pi(\theta)=1$. 这就是 Bayes 假设.

3. 尺度参数的无信息先验

设总体 X 的密度具有形式$\frac{1}{\sigma}p\left(\frac{x}{\sigma}\right)$，其中 σ 称为尺度参数，参数空间为$(0,\infty)=\mathbf{R}^+$. 这类密度组成尺度参数族.

设想让 X 改变比例尺度得 $Y=cX(c>0)$，类似地，也让 σ 同步变化，即 $\eta=c\sigma$. 不难看出，Y 的密度函数为$\frac{1}{\eta}p\left(\frac{y}{\eta}\right)$，仍属尺度参数族. 因此 σ 的无信息先验分布 $\pi(\sigma)$与 η 的无信息先验分布 $\pi^*(\eta)$应相同，即

$$\pi(\tau) = \pi^*(\tau)$$

另一方面，由变换 $\eta=c\sigma$ 可以算得 η 的无信息先验分布

$$\pi^*(\eta) = \frac{1}{c}\pi\left(\frac{\eta}{c}\right)$$

比较上面两式可得 $\pi(\eta)=\frac{1}{c}\pi\left(\frac{\eta}{c}\right)$，取 $\eta=c$，则有 $\pi(c)=\frac{1}{c}\pi(1)$. 为方便计，令 $\pi(1)=1$，这样就得到 σ 的无信息先验分布

$$\pi(\sigma) = \frac{1}{\sigma}, \quad \sigma > 0$$

这表明：尺度参数在按比例变换下保持不变的无信息先验分布是 $\pi(\sigma)=\sigma^{-1}$.

这仍是一个不正常先验分布. 但可成为广义先验分布. 譬如它与单参数指数分布 $p(x/\sigma)=\frac{1}{\sigma}\exp\{-x/\sigma\}$结合而得到的后验为 $\pi(\sigma|x)=k\sigma^{-2}\exp\{-x/\sigma\},\sigma>0$. 这仍是倒 Gamma 分布 $IG(1,x)$. 它是一个正常密度函数. 故上述无信息先验分布是尺度参数 σ 的广义先验分布.

4. Jeffreys 先验分布

设 $x=(x_1,\cdots,x_n)$是来自密度函数 $p(x|\theta)$的一个样本，其中 $\boldsymbol{\theta}=(\theta_1,\cdots,\theta_p)$是 p 维参数向量. 在对 $\boldsymbol{\theta}$ 无任何先验信息可用时，Jeffreys (1961)利用变换群和 Harr 测度导出 $\boldsymbol{\theta}$ 的无信息先验分布可用 Fisher 信息阵的行列式的平方根表示. 这种无信息先验分布常称为 Jeffreys 先验分布，其寻求步骤如下：

(i) 写出样本的对数似然函数 $l(\theta|x)=\sum_{i=1}^{n}\ln p(x_i|\theta)$；

(ii) 算出参数 $\boldsymbol{\theta}$ 的 Fisher 信息阵

$$I(\theta) = \mathrm{E}_{x|\theta}\left(-\frac{\partial^2 l}{\partial\theta_i\partial\theta_j}\right)_{i,j=1,\cdots,p}$$

在单参数场合，$I(\theta)=\mathrm{E}_{x|\theta}\left(-\frac{\partial^2 l}{\partial\theta^2}\right)$；

(iii)θ 的无信息先验密度函数为 $\pi(\theta)=[\det I(\theta)]^{1/2}$. 在单参数场合，$\pi(\theta)=[I(\theta)]^{1/2}$.

例 5.37 设 $X=(x_1,\cdots,x_n)$ 是来自正态总体 $N(\mu,\sigma^2)$ 的一个样本，现求参数向量 (μ,σ) 的 Jeffreys 先验.

容易写出其对数似然函数

$$l(\mu,\sigma) = -\frac{1}{2}\ln(2\pi) - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

其 Fisher 信息阵为

$$I(\mu,\sigma)=\begin{pmatrix}\mathrm{E}\left(-\frac{\partial^2 l}{\partial\mu^2}\right) & \mathrm{E}\left(-\frac{\partial^2 l}{\partial\mu\partial\sigma}\right)\\ \mathrm{E}\left(-\frac{\partial^2 l}{\partial\mu\partial\sigma}\right) & \mathrm{E}\left(-\frac{\partial^2 l}{\partial\sigma^2}\right)\end{pmatrix}$$

$$=\begin{pmatrix}n/\sigma^2 & 0\\ 0 & 2n/\sigma^2\end{pmatrix}$$

$$\det I(\mu,\sigma) = 2n^2\sigma^{-4}$$

所以(μ,σ)的 Jeffreys 先验为

$$\pi(\mu,\sigma) \propto \sigma^{-2}$$

它的几个特例是

(1) 当 σ 已知时，$I(\mu)=\mathrm{E}\left(\frac{\partial^2 l}{\partial\mu^2}\right)=n/\sigma^2$. 故 $\pi(\mu)=1,\mu\in\mathbf{R}$；

(2) 当 μ 已知时，$I(\sigma)=\mathrm{E}\left(\frac{\partial^2 l}{\partial\sigma^2}\right)=2n/\sigma^2$. 故 $\pi(\sigma)=1/\sigma,\sigma\in\mathbf{R}^+$；

(3) 当 μ 与 σ 相互独立时，$\pi(\mu,\sigma)=1/\sigma,\mu\in\mathbf{R},\sigma\in\mathbf{R}^-$.

可见 Jeffreys 先验分布表明：μ 与 σ 的无信息先验分布是不独立的. 在(μ,σ)的联合无信息先验分布的两种形式(σ^{-1}与 σ^{-2})中，Jeffreys 最终推荐的是 $\pi(\mu,\sigma)=1/\sigma$. 实际的使用情况看，多数人愿意采用 Jeffreys 的最终推荐形式.

例 5.38 设 θ 为成功概率，则在 n 次独立试验中成功次数 X 服从

二项分布,即

$$P(X=x)=\binom{n}{x}\theta^x(1-\theta)^{n-x},\quad x=0,1,\cdots,n$$

其对数似然函数和 Fisher 信息分别为

$$l=x\ln\theta+(n-x)\ln(1-\theta)+\ln\binom{n}{x}$$

$$I(\theta)=\mathrm{E}\left(-\frac{\partial^2 l}{\partial\theta^2}\right)=\frac{n}{\theta}+\frac{n}{1-\theta}$$

所以,在二项分布场合,成功概率 θ 的 Jeffreys 先验分布为

$$\pi(\theta)\propto\theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}},\quad \theta\in(0,1)$$

关于成功概率 θ 的无信息先验分布,至今至少已有以下四种:

$\pi_1(\theta)=1$ -Bayes(1763)和 Laplace(1812)采用过.

$\pi_2(\theta)=\theta^{-1}(1-\theta)^{-1}$ -Novick 和 Hall(1965)导出.

$\pi_3(\theta)=\theta^{-1/2}(1-\theta)^{-1/2}$ -Jeffreys(1968)导出.

$\pi_4(\theta)=\theta^{\theta}(1-\theta)^{1-\theta}$ -Zellner(1977)导出.

所有这四种无信息先验都是合理的,因为他们各从一个侧面提出自己的合理要求,然后再推导出来的. 其中 π_2 是不正常密度,而 π_1 是正常密度,而 π_3,π_4 经过正则化后亦可成为正常密度. 但它们与二项分布结合都可得正常后验密度,故它们亦都是广义先验分布.

一般说来,无信息先验不是唯一的,但它们对 Bayes 统计推断的结果的影响都是很小的,很少对结果产生重大影响. 所以任何无信息先验分布都可采用. 当今无论在统计理论研究和应用研究中,采用无信息先验分布愈来愈多. 连经典统计学者也认为无信息先验是"客观"的,可以接受的,这也是近几十年来 Bayes 学派研究中最成功的部分.

§ 5.5.4 多层先验分布

当所给定先验分布中超参数难于确定时,可以对超参数再给出一个先验,第二个先验称为超先验. 由先验和超先验决定的一个新先验就称为多层先验. 下面的例子可以帮助我们理解多层先验的想法和做法.

例 5.39 设对某产品的不合格品率 θ 了解甚少,只知道它比较

小，现要确定 θ 的先验分布，决策人经过反复思考，最后把他引导到多层先验上去，他的思路是这样的.

(1) 开始他用区间(0,1)上的均匀分布 $U(0,1)$ 作为 θ 的先验分布.

(2) 后觉得不妥，因为此产品的不合格品率 θ 比较小，不会超过 0.5，于是他改用区间(0,0.5)上的均匀分布 $U(0,0.5)$ 作为 θ 的先验分布.

(3) 在一次业务会上不少人对上限 0.5 提出各种意见，有人问："为什么不把上限定为 0.4 呢?"他讲不清楚. 有人建议："上限很可能是 0.1"，他也无把握，但这些问题促使他思考，最后他把自己的思想理顺了，提出如下看法：θ 的先验为 $U(0,\lambda)$，其中 λ 是超参数，要确切地定出 λ 是困难的，但预测出它的区间倒是有把握的，根据大家建议，他认为 λ 是在区间(0.1,0.5)上的均匀分布 $U(0.1,0.5)$. 这后一个分布被称为超先验，决策人的这种归纳获得大家的赞许.

(4) 最后决定的先验是什么呢? 根据决策人的归纳，可以叙述为如下两点.

a. θ 的先验为 $\pi_1(\theta|\lambda)=U(0,\lambda)$.

b. λ 的超先验为 $\pi_2(\lambda)=U(0.1,0.5)$.

于是用边际分布计算公式，可得 θ 的先验为

$$\pi(\theta)=\int_{\Lambda}\pi_1(\theta|\lambda)\pi_2(\lambda)\mathrm{d}\lambda$$

其中 Λ 是超参数 λ 的取值范围. 在这个例子中

$$\pi(\theta)=\frac{1}{0.5-0.1}\int_{0.1}^{0.5}\lambda^{-1}I_{(0,\lambda)}(\theta)\mathrm{d}\lambda$$

其中 $I_A(\theta)=1$，当 $\theta\in A$，否则 $I_A(\theta)=0$，分几种情况计算上述积分.

当 $0<\theta<0.1$ 时，

$$\pi(\theta)=\frac{1}{0.4}\int_{0.1}^{0.5}\lambda^{-1}\mathrm{d}\lambda=2.5\ln 5=4.0236$$

当 $0.1\leqslant\theta<0.5$ 时，

$$\pi(\theta)=2.5\int_{\theta}^{0.5}\lambda^{-1}\mathrm{d}\lambda=2.5[\ln 0.5-\ln\theta]=-1.7329-2.5\ln\theta$$

当 $0.5\leqslant\theta$ 时，$\pi(\theta)=0$.

综合上述，最后得到的多层先验（见图 5.9）为

$$\pi(\theta)=\begin{cases}4.0236, & 0\leqslant\theta<0.1\\ -1.729-2.5\ln\theta, & 0.1\leqslant\theta<0.5\\ 0, & 0.5\leqslant\theta<1\end{cases}$$

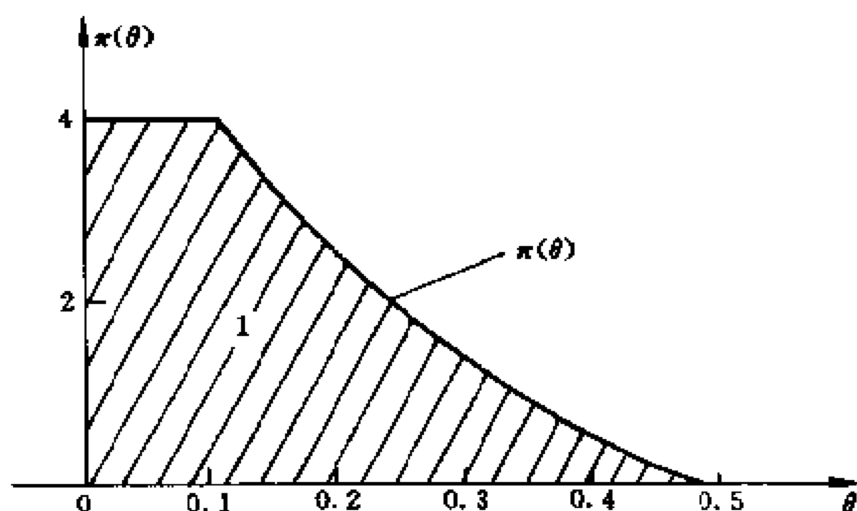


图 5.9 θ 的多层先验

这是一个正常先验. 容易验证 $\int_0^1\pi(\theta)\mathrm{d}\theta=1$.

从上述例子可以看出一般多层先验的确定方法：

首先对未知参数 θ 给出一个形式已知的密度函数作为先验分布，即 $\theta\sim\pi_1(\theta|\lambda)$，其中 λ 是超参数. Λ 是其取值范围.

第二步是对超参数 λ 在给出一个超先验 $\pi_2(\lambda)$，

由此可得多层先验的一般表现形式

$$\pi(\theta)=\int_\Lambda\pi_1(\theta|\lambda)\pi_2(\lambda)\mathrm{d}\lambda$$

而任一个 Bayes 分析都是对 $\pi(\theta)$ 进行的.

应该说明在理论上没有限制多层先验只分两步，可以是三步或更多步，但在实际应用中多于两步的先验是罕见的，对第二步先验 $\pi_2(\lambda)$ 用主观概率或用历史数据给出是有困难的，因为 λ 常是不能观察的，甚至连间接观察都是难于进行的. 很多人用无信息先验作为第二步试验

是一种好的策略. 因为第二步先验即使决定有偏差,而导致错误结果的危险性更小一些,相对说来,第一步先验更为重要些.

多层先验常常是在这样一种场合使用,当一步给出先验 $\pi(\theta)$ 没有把握时,那用两步先验要比硬用一步先验所冒风险要小一些.

例 5.40　设有 m 位优秀学生参加智力测验,第 i 位学生的考分 X_i 可看作来自正态分布 $N(\theta_i,\sigma^2)$ 的一个样本,其中 θ_i 是第 i 位学生的能力,方差 σ^2 是已知的. 这个说明对 $i=1,2,\cdots,m$ 都适用. 也就是说,$X_1,X_2,\cdots,X_m$ 是来自方差相同,均值(能力)各不相同的一个正态样本. 现要寻求 m 位学生的能力 $\theta_1,\theta_2,\cdots,\theta_m$ 的联合先验分布. 下面用多层先验方法给出它.

按经验,这 m 个优秀学生的能力 $\theta=(\theta_1,\theta_2,\cdots,\theta_m)$ 可以看作是来自某个正态分布 $N(\mu_\pi,\sigma_\pi^2)$ 的一个样本. 因此第一层先验可选为

$$\pi_1(\theta|\lambda) = \frac{1}{(\sqrt{2\pi}\sigma_\pi)^m}\exp\left\{-\frac{1}{2\sigma_\pi^2}\sum_{i=1}^m(\theta_i-\mu_\pi)^2\right\}$$

其中 $\lambda=(\mu_\pi,\sigma_\pi^2)$ 是两个超参数,对超参数 μ_π 依过去经验可认为是 100,即 $\mu_\pi=100$,而要确定 σ_π^2 具体值无多大把握,只能说一个大概. 真实能力的方差 σ_π^2 服从倒 Gamma 分布 $IG(\alpha,\lambda)$,该分布均值为 200,方差为 100,由倒 Gamma 分布的均值与方差可以列出矩的方程

$$\begin{cases}\dfrac{\lambda}{\alpha-1} = 200\\[2ex]\dfrac{\lambda^2}{(\alpha-1)^2(\alpha-2)} = 100^2\end{cases}$$

解之可得,$\alpha=6,\lambda=1000$. 于是可写出第二层先验

$$\pi_2(\sigma_\pi^2) = \frac{1}{\Gamma(6)}\cdot\frac{1000^6}{(\sigma_\pi^2)^{6+1}}\mathrm{e}^{-1000/\sigma_\pi^2},\quad \sigma_\pi^2>0$$

综合上述两层先验,可得 θ 的先验为

$$\begin{aligned}\pi(\theta) &= \int_0^\infty \pi_1(\theta|\sigma_\pi^2)\,\pi_2(\sigma_\pi^2)\,\mathrm{d}\sigma_\pi^2\\ &= \int_0^\infty \frac{1}{(2\pi\sigma_\pi^2)^{m/2}}\exp\left\{-\frac{1}{2\sigma_\pi^2}\sum_{i=1}^m(\theta_i-100)^2\right\}\cdot\end{aligned}$$

$$\frac{10^{18}}{120(\sigma_\pi^2)^7}e^{-1000/\sigma_\pi^2}d\sigma_\pi^2$$

$$= \frac{10^{18}}{120(2\pi)^{m/2}} \int_0^\infty (\sigma_\pi^2)^{-(7+\frac{m}{2})} \cdot$$

$$\exp\left\{-\frac{1}{2\sigma_\pi^2}\left[1000 + \frac{1}{2}\sum_{i=1}^m(\theta_i - 100)^2\right]\right\}d\sigma_\pi^2$$

$$= \frac{10^{18}}{120(2\pi)^{m/2}} \cdot$$

$$\frac{\Gamma\left(6+\frac{m}{2}\right)}{\left[1000 + \frac{1}{2}\sum_{i=1}^m(\theta_i - 100)^2\right]^{6+m/2}}$$

$$= \frac{\Gamma\left(\frac{12+m}{2}\right)}{\Gamma(6)(2000\pi)^{m/2}}\left[1 + \frac{1}{2000}\sum_{i=1}^m(\theta_i - 100)^2\right]^{-\frac{12+m}{2}}$$

这是什么分布呢？一般的 m 维 t 分布的密度函数为

$$f(\boldsymbol{x}|\alpha,\mu,\Sigma)$$

$$= \frac{\Gamma\left(\frac{\alpha+m}{2}\right)}{(\det\boldsymbol{\Sigma})^{1/2}(\alpha\pi)^{m/2}\Gamma\left(\frac{\alpha}{2}\right)}\left[1 + \frac{1}{\alpha}(\boldsymbol{x}-\mu)'\Sigma^{-1}(\boldsymbol{x}-\mu)\right]^{-\frac{\alpha+m}{2}}$$

其中参数 α 是其自由度，向量 μ 为位置参数向量，矩阵 $\boldsymbol{\Sigma}$ 是其尺度矩阵. 并且 $\boldsymbol{x}\in\mathbf{R}^m$，$\alpha>0$，$\mu\in\mathbf{R}^m$，$\boldsymbol{\Sigma}$ 是 $p\times p$ 阶正定矩阵，对照上述 m 维 t 分布，可以看出，这里所得到的先验密度 $\pi(\theta)$ 是 $\alpha=12$，$\boldsymbol{\mu}=(100,\cdots,100)'$，$\Sigma=\left(\frac{500}{3}\right)I$ 的 m 维 t 分布的密度函数. 这种先验分布要一步依据主观信念和历史数据获得是很困难的，甚至很难想到. 多层先验的这个优点特别明显.

§ 5.5.5 可信域

Bayes 分析中另一个常见的推断形式是确定参数的可信域，它是利用样本 x 和获得的后验分布 $\pi(\theta|x)$ 在参数空间 Θ 中寻找一个区域 C_x，使得其后验概率 $P_{\theta|x}(\theta\in C_x)$ 尽可能地大，而其容量 $\mathrm{vol}(C_x)$ 尽可能

地小，其中 vol 是 volume(容量)的前几个字母. 这些要求可用如下损失函数表述出来.

$$L(\theta,C) = m_1\mathrm{vol}(C) + m_2[1 - I_C(\theta)]$$

其中 m_1 与 m_2 是两个权数，皆为正. 它表示决策者的重视程度. I_C 是区域 C 的示性函数. 其后验风险就是

$$\mathrm{E}_{\theta|x}L(\theta,C_x) = m_1\mathrm{E}_{\theta|x}[\mathrm{vol}(C_x)] + m_2[1 - P_{\theta|x}(\theta \in C_x)]$$

上式中第一项是与平均容量(精度)成正比的量，第二项是与 $\theta \notin C_x$ 的后验概率(可靠度)成正比的量. 要使后验风险最小就要求上述两项都很小，这是很难做到的. 因为实践告诉我们，这两项要求是矛盾的，一个常用的折衷方案是在后验概率 $P_{\theta|x}(\theta \in C_x)$ 达到一定要求下，尽量使平均容量尽可能地小. 这就产生如下最大后验密度(Highest Posterier Density，简称 HPD)可信域的概念.

定义 5.13 设参数 θ 的后验密度函数为 $\pi(\theta|x)$，对给定的样本 x 和概率 $1-\alpha(0<\alpha<1)$，若在参数空间上存在这样的区域 C_x，并满足下列两个条件

(i) $P_{\theta|x}(C_x) \geqslant 1-\alpha$;

(ii) 任给 $\theta_1 \in C_x$ 和 $\theta_2 \notin C_x$，总有

$$\pi(\theta_1|x) \geqslant \pi(\theta_2|x)$$

则称 C_x 是可信水平为 $1-\alpha$ 的最大后验密度可信域. 简称为 $(1-\alpha)$ HPD 可信域. 在 θ 为一维场合，若 C_x 是一个区间，则又称 C_x 为 $(1-\alpha)$ HPD 可信区间.

从上述定义可见，寻找 $(1-\alpha)$HPD 可信域就是要在 Θ 中寻找形如

$$C_x = \{\theta : \pi(\theta|x) > k(\alpha), \theta \in \Theta\}$$

的区域，其中 $k(\alpha)$ 是使

$$P_{\theta|x}(C_x) \geqslant 1 - \alpha$$

成立的最大常数. 在连续场合，可能有 $P_{\theta|x}(C_x)=1-\alpha$. 在离散场合，可能是 $P_{\theta|x}(C_x)\geqslant 1-\alpha$. 当 θ 为一维参数时，后验分布呈单峰状可得 HPD 可信区间(图 5.10(a))，在离散场合，可能得到几个不相连结的区间组成的 HPD 可信域(图 5.10(b))

例 5.41 设 $X_1,X_2,\cdots,X_n$ 是来自正态总体 $N(\theta,\sigma^2)$ 的一个样本，

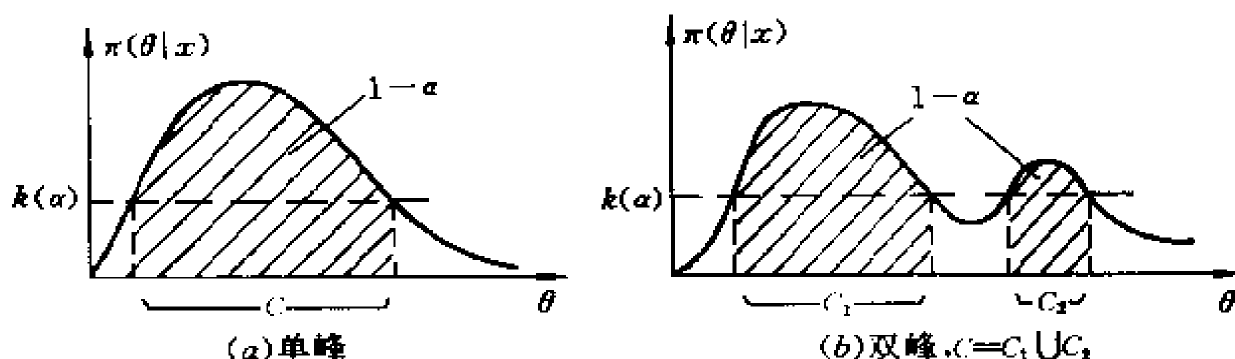


图 5.10 可信区间与可信域

其中 σ 已知. 正态均值 θ 的先验分布是 $N(\mu,\tau^2)$，其中 μ 与 τ^2 已知. 在例 5.26 中已求得 θ 的后验分布为 $N(\mu_1,\sigma_1^2)$，其中

$$\mu_1=\frac{\sigma_0^{-2}\mu+\tau^{-2}\overline{x}}{\sigma_0^{-2}+\tau^{-2}},\quad \sigma_1^2=(\sigma_0^{-2}+\tau^{-2})^{-1}$$

$\overline{x}$ 为样本均值，$\sigma_0^2=\sigma^2/n$. 如今要求正态均值 θ 的 $(1-\alpha)$ 可信区间.

由于后验分布 $N(\mu_1,\sigma_1^2)$ 的对称性，θ 的 $(1-\alpha)$HPD 可信区间容易获得，它是 $[\mu_1-\sigma_1u_{1-\alpha/2},\mu_1+\sigma_1u_{1-\alpha/2}]$，其中 $u_{1-\alpha/2}$ 是标准正态分布的 $1-\alpha/2$ 分位数.

在儿童智商测验(见例 5.31)中，$X\sim N(\theta,100)$，$\theta\sim N(100,225)$. 在仅取一个样本 $(n=1)$ 情况下，算得一儿童智商 θ 的后验分布为 $N(\mu_1,\sigma_1^2)$，其中

$$\mu_1=\frac{400+9x}{13},\quad \sigma_1^2=69.23=(8.32)^2$$

该儿童在一次智力测验中得 $x=115$ 分. 在平方损失函数下 θ 的贝叶斯估计为 $\hat{\theta}=110.38$. 如今来求 θ 的 0.95 可信区间. 由于 $u_{0.975}=1.96$，故可算得

$$\mu_1-\sigma_1u_{0.975}=110.38-8.32\times1.96=94.07$$

$$\mu_1+\sigma_1u_{0.975}=110.38+8.32\times1.96=126.69$$

该儿童智商 θ 的 0.95 可信区间为 $[94.07,126.69]$，该区间长为 32.62.

例 5.42 设 $x_1,\cdots,x_n$ 是来自正态总体 $N(\theta,\sigma^2)$ 的一个样本，若取

Jeffreys 先验分布，即

$$\pi(\theta,\sigma) = \frac{1}{\sigma^2}, \quad (\theta,\sigma^2) \in \mathbf{R} \times \mathbf{R}^+$$

则其后验分布为

$$\pi(\theta,\sigma^2|\bar{x},s) \propto \sigma^{-(n+2)}\exp\left\{-\frac{s^2 + n(\bar{x}-\mu)}{2\sigma^2}\right\}$$

其中 $\bar{x} = \sum_{i=1}^n x_i/n, s^2 = \sum_{i=1}^n (x_i - \bar{x})^2/(n-1)$. 现来分别寻找 θ 和 σ^2 的 HPD 可信域. 首先我们求 θ 的边际分布

$$\pi(\theta|\bar{x},s) \propto \int_0^\infty (\sigma^2)^{-\frac{n}{2}-1}\exp\left\{-\frac{[s^2 + n(\theta-\bar{x})]/2}{\sigma^2}\right\}\mathrm{d}\sigma^2$$

上述被积函数是倒 Gamma 分布 $IG(\alpha,\lambda)$, $\alpha=\frac{n}{2}$, $\lambda=[s^2+n(\theta-\bar{x})^2]/2$ 的核. 利用正则性，可得

$$\pi(\theta|\bar{x},s) \propto [s^2 + n(\theta - \bar{x})^2]^{-\frac{n}{2}} \propto \left[1 + \left(\frac{\theta - \bar{x}}{s/\sqrt{n}}\right)^2\right]^{-\frac{(n-1)-1}{2}}$$

这是如下一般 t 分布（习题 1.18）的核，自由度为 ν、位置参数为 $\mu\in\mathbf{R}$，尺度参数为 $\tau\in\mathbf{R}^+$ 的一般 t 分布的密度函数为

$$p(y|\nu,\mu,\tau) = \frac{\Gamma\left(\frac{\nu+2}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}\tau}\left[1 + \frac{1}{\nu}\left(\frac{y-\mu}{\tau}\right)^2\right]^{-\frac{\nu+1}{2}}, \quad y \in \mathbf{R}$$

当 $\mu=0,\tau=1$ 即得 $t(\nu)$ 变量. 这里若取

$$\nu = \frac{n-1}{2}, \quad \mu = \bar{x}, \quad \tau = \sqrt{\frac{n-1}{n}}s$$

即可看出：上述 θ 的边际密度是一般 t 分布. 从而在给定 $\bar{x}$ 和 s 条件下，

$$\frac{\sqrt{n}(\theta - \bar{x})}{s} \sim t(n-1)$$

它的 $(1-\alpha)$HPD 可信区间为

$$C_x = \left[\bar{x} - t_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{1-\alpha/2}\frac{s}{\sqrt{n}}\right]$$

其中 $t_{1-\alpha/2}$是 $t(n-1)$的 $1-\alpha/2$ 分位数. 这与经典结果完全一致. 这种

现象在使用无信息先验分布时常会出现，这可能是“无先验信息”之故.

其次，我们来寻求 σ^2 的边际分布

$$\pi(\sigma^2|\bar{x},s) \propto \sigma^{-(n+2)}e^{-s^2/2\sigma^2}\int_{-\infty}^{\infty}\exp\left\{-\frac{(\theta-\bar{x})^2}{2\sigma^2/n}\right\}d\theta$$

这是倒 Gamma 分布 $IG(\alpha,\lambda)$，$\alpha=(n-1)/2$，$\lambda=s^2/2$ 的核. 从而有

$$\frac{s^2}{\sigma^2}\sim Ga\left(\frac{n-1}{2},\frac{1}{2}\right)=\chi^2(n-1)$$

这与经典结果完全一致，由于它是偏正态分布，寻求 $(1-\alpha)$HPD 可信区间要借助计算机才能实现.

当后验密度 $\pi(\theta|x)$ 为一维单峰连续函数时，寻求 θ 的 $(1-\alpha)$HPD 可信区间的数值计算一般可按下述思路进行.

1° 对给定的 k，建立子程序：以方程

$$\pi(\theta|x)=k$$

可解得 $\theta_1(k)$ 和 $\theta_2(k)$，使

$$C(k)=\{\theta;\pi(\theta|x)\geqslant k\}=[\theta_1(k),\theta_2(k)]$$

2° 建立子程序，用来计算概率

$$P_{\theta|x}(C(k)|x)=\int_{C(k)}\pi(\theta|x)d\theta$$

3° 对给定的 k，

若 $P_{\theta|x}(C(k)|x)\approx 1-\alpha$，则 $C(k)$ 即为所求的 HPD 可信区间；

若 $P_{\theta|x}(C(k)|x)>1-\alpha$，则增大 k，转入 1° 与 2°；

若 $P_{\theta|x}(C(k)|x)<1-\alpha$，则减小 k，转入 1° 与 2°.

例 5.43　在例 5.32 中已确定彩电平均寿命 θ 的后验分布为倒伽玛分布 $IG(1.956, 42\,868)$. 现求 θ 的可信水平为 0.90 的最大后验密度(HPD)可信区间.

解：为简单起见，这里的 1.956 用近似数 2 代替，于是 θ 的后验密度为

$$\pi(\theta|t)=\lambda^2\theta^{-3}e^{-\lambda/\theta},\quad \theta>0$$

其中 t 表示截尾样本，$\lambda=42\,868$. 它的分布函数可以表示为

$$F(\theta'|t)=P(\theta\leqslant\theta'|t)=\int_0^{\theta'}\lambda^2\theta^{-3}e^{-\lambda/\theta}d\theta=\left(1+\frac{\lambda}{\theta'}\right)e^{-\lambda/\theta'},\quad \theta'>0$$

这就为计算落入可信区间的概率提供方便.

另外,此后验密度是单峰函数,其众数 $\text{Mod}(\theta|t)=\lambda/3=14\ 289$.这就告诉我们,$\theta$ 的 HPD 可信区间的两个端点分别在此众数两侧,在这一点(众数)上的后验密度函数值为

$$\pi(\theta|t)=\lambda^2\left(\frac{3}{\lambda}\right)^3 e^{-3}=\frac{27}{e^3\times 42868}=0.000\ 031\ 358$$

这个数过小,在以下的计算中我们用 $\lambda\pi(\theta|t)$ 来代替 $\pi(\theta|t)$,这并不会影响我们寻求 HPD 可信区间,其中

$$\lambda\pi(\theta|t)=\left(\frac{\lambda}{\theta}\right)^3\exp\left\{-\left(\frac{\lambda}{\theta}\right)\right\}$$

我们按寻求 HPD 可信区间的程序 $1^\circ\sim 3^\circ$ 进行,经过四轮计算就获得 θ 的 0.90 的 HPD 可信区间(4.735,81 189),即

$$P(4.735<\theta<81\ 189|t)=0.90$$

具体计算如下:在第一轮中,我们首先取 $\theta_U^{(1)}=42\ 868$(由于它大于 $\text{Mod}(\theta|t)$,故一定可当作可信区间上限),并可算得

$$\lambda\pi(\theta_U^{(1)}|t)=0.367\ 879$$

然后在计算机上搜索,发现当 $\theta_L^{(1)}=6.387$ 时,即 $\lambda/\theta_L^{(1)}=6.7115$ 时,有

$$\lambda\pi(\theta_L^{(1)}|t)=0.367\ 877$$

这时可认为 $\lambda\pi(\theta_U^{(1)}|t)=\lambda\pi(\theta_L^{(1)}|t)=0.3679$,其可信区间 $(\theta_L^{(1)},\theta_U^{(1)})$ 的概率为

$$\begin{aligned}P(\theta_L^{(1)}\leqslant\theta\leqslant\theta_U^{(1)}|t)&=P(\theta\leqslant\theta_U^{(1)}|t)-P(\theta<\theta_L^{(1)}|t)\\&=0.735\ 76-0.009\ 38=0.726\ 38\end{aligned}$$

此概率比 0.90 要小,故还需扩大区间.

在第二轮中,我们取 $\theta_U^{(2)}=85\ 736$,这时 $\lambda/\theta_U^{(2)}=0.5$,

$$\lambda\pi(\theta_U^{(2)}|t)=0.075\ 816$$

然后在计算机上搜索,发现当 $\theta_L^{(2)}=4.632$,即 $\lambda/\theta_L^{(2)}=9.255$ 时,有

$$\lambda\pi(\theta_L^{(2)}|t)=0.075\ 811$$

可以认为 $\lambda\pi(\theta_U^{(2)}|t)=\lambda\pi(\theta_L^{(2)}|t)=0.0758$. 其可信区间的概率为

$$P(\theta_L^{(2)}\leqslant\theta\leqslant\theta_U^{(2)}|t)=0.909\ 800-0.000\ 981=0.908\ 819$$

此概率又比 0.90 大一点,还要缩小区间.

我们把上述搜索过程列于表 5.6 中，第三、四轮也列出，从最后的结果可见，(4.735，81 189)是 θ 的 0.90HPD 可信区间.

表 5.6 HPD 可信区间的搜索过程

θ_0	λ/θ_0	$\lambda\pi(\theta_0\|t)=\left(\frac{\lambda}{\theta_0}\right)^3 e^{-\lambda/\theta_0}$	$P(\theta\leqslant\theta_0)=\left(1+\frac{\lambda}{\theta_0}\right)e^{-\lambda/\theta_0}$	$P(\theta_L\leqslant\theta\leqslant\theta_U/t)$
$\theta_U^{(1)}=12\,868$	1	0.367 879	0.735 759	
$\theta_L^{(1)}=6.387$	6.71	0.36 775	0.009 383	0.726 376
$\theta_U^{(2)}=85\,736$	0.5	0.75 816	0.909 800	
$\theta_L^{(2)}=4.632$	9.255	0.075 811	0.000 981	0.908 819
$\theta_U^{(3)}=80\,883$	0.53	0.087 630	0.900 560	
$\theta_L^{(3)}=4.742$	9.039	0.087 654	0.001 191	0.898 375
$\theta_U^{(4)}=81\,189$	0.528	0.086 815	0.901 189	
$\theta_L^{(4)}=4.735$	9.053	0.086 838	0.001 177	0.900 012

在寻求参数 θ 的可信域中，其计算虽比经典方法简单一些，但要寻求 HPD 可信域会遇到一些困难. 特别当 HPD 可信域不能联成一片(如图 5.10(b)所示)时，可放弃 HPD 准则而用相连的可信域. 譬如，在 θ 为单参数时，代替$(1-\alpha)$HPD 可信域可用等尾的可信区间$[C_L(x), C_U(x)]$，使得

$$P_{\theta|x}(\theta < C_L(x)) = P_{\theta|x}(\theta > C_U(x)) = \alpha/2$$

Berger 指出[2]：当后验密度函数出现多峰时，常常是由于先验分布与观察值不一致引起的. 问题可能是先验分布选择不当，或是抽样分布选择不当引起的. 认识和研究此种抵触信息往往是重要的. 而在使用共轭先验分布时，若共轭先验分布是单峰的，将导致后验分布也是单峰的. 这时若有抵触信息存在就被掩盖了，故要慎重对待和使用共轭先验分布.

由于实际需要，常要寻求一维参数 θ 的单侧可信限这也得放弃 HPD 准则，把满足

$$P_{\theta|x}(\theta \geqslant C_L(x)) \geqslant 1-\alpha$$

的 $C_L(x)$ 称为 θ 的 $(1-\alpha)$ 可信下限，满足

$$P_{\theta|x}(\theta \leqslant C_U(x)) \geqslant 1-\alpha$$

的 $C_U(x)$ 称为 θ 的 $(1-\alpha)$ 可信上限，在后验分布 $\pi(\theta|x)$ 为连续时，上述两个后验概率可以达到 $1-\alpha$.

例 5.44　在例 5.32 中已确定彩电平均寿命 θ 的后验分布为倒 Gamma 分布 $IG(\alpha,\lambda)$，其中 $\alpha=1.956$，$\lambda=42\,868$. 这是在截尾样本 $\boldsymbol{t}=(t_1,\cdots,t_r)$ 下得到的 θ 的分布，即

$$\theta|t \sim IG(\alpha,\lambda)$$

由此可得

$$2\lambda/\theta|t \sim Ga\left(\alpha,\frac{1}{2}\right)=\chi^2(2\alpha)$$

于是 θ 的 90% 可信下限为

$$C_L(t)=2\lambda/\chi^2_{0.90}(2\alpha)$$

这里自由度 $f=2\alpha=3.912$. 从 χ^2 分位数表上可查得 $\chi^2_{0.90}(3)=6.251$，$\chi^2_{0.90}(4)=7.779$. 用线性内插法可获得近似值 $\chi^2_{0.90}(3.912)\approx 7.645$. 于是 θ 的 90% 可信下限为

$$C_L(t)=\frac{2\times 42\,868}{7.645}=11\,215(\text{h})$$

即在八十年代中期我国彩电平均寿命的 90% 可信下限为 11 215 小时.

最后，我们指出，当 θ 是 p 维参数向量 $(p\geqslant 2)$ 且后验分布较为复杂时，可用第六章叙述的数值模拟方法去获得可信域，也可用 p 元正态分布近似，即

$$\pi(\boldsymbol{\theta}|x) \approx N_p(\mathrm{E}(\boldsymbol{\theta}|x),\mathrm{Var}(\boldsymbol{\theta}|x))$$

其中 $\mathrm{E}(\boldsymbol{\theta}|x)$ 和 $\mathrm{Var}(\boldsymbol{\theta}|x)$ 分别为后验分布的均值向量与方差协方差阵. 由此可导出近似的可信域.

$$C_x=\{\boldsymbol{\theta}:(\boldsymbol{\theta}-\mathrm{E}(\boldsymbol{\theta}|x))'[\mathrm{Var}(\boldsymbol{\theta}|x)]^{-1}(\boldsymbol{\theta}-\mathrm{E}(\boldsymbol{\theta}|x))\leqslant \chi^2_\alpha(p)\}$$

其中 $\chi^2_\alpha(p)$ 为自由度为 p 的 χ^2 分布 α 分位数. 但这仅在大样本场合才可使用.

参考文献

1 Wald A. Statistical decision functions. New York: John Wiley-Sons, 1950. 中译本:王福保译,统计决策函数,上海:上海教育出版社,1963

2 Berger J O. Statistical decision theory and Bayesian analysis . 2nd ed. New York: Springer-Verlag, 1985. 中译本:贾乃光译. 统计决策理论和贝叶斯分析. 北京:中国统计出版社,1998

3 Gelman A, Carlin J B, Stern H S, Bubin D B. Bayesian data analysis. New York: Chapman-Hall, 1995

4 Press S J. Bayesian Statistics. New York: John Wiley-Sons, 1989. 中译本:廖文,陈安贵译. 贝叶斯统计学. 北京:中国统计出版社,1992

5 张尧庭,陈汉峰. 贝叶斯统计推断. 北京:科学出版社,1991

6 张雪野,茆诗松. 经营决策方法. 上海:华东师范大学出版社,1996

7 Mckinsey. Introduction to the theory of games. 中译本:高鸿勋等译. 博奕论导引. 北京:人民教育出版社,1960

8 Cox D R, Hinkley D V. Theoretical Statistics. London: Chapman-Hall, 1974

9 Stein C. Imadmissibility of the usual estimator for the mean of a multivariate normal distribution. In Pro: 3rd Berkely Symp, Math. Statist. Probab. Berkeley: Univ of California Press, 1955, 1:197～206

10 James W, Stein C. Estimations with quadratic loss. In Pro: Fourth Berkely Symp. Math. Statist . Probab. Berkeley: Univ of California Press, 1961, 1:361～380

11 Baranchick A J. A family of minimax estimators of the mean of a multivariate normal distribution. Ann Math Statist, 1970, 41:642～645

12　Cellier D, Fourdrinier D, Robert C. Robust shrinkage estimators of the location parameter for clitically symmetric distribution. J Multivariate Anal, 1989, 29: 39～52

13　Christian P R. The Bayesian choice-A decision theoretic motivation. New York: Springer-Verlag, 1994

14　成平，陈希孺，陈桂景，吴传义. 参数估计. 上海：上海科学技术出版社，1985

15　Pratt J W. Discussion of A. Birnbaum's "On the foundations of statistical inference." J Amer Statist Assoc, 1962, 57: 269～326

16　Jeffreys H. Theory of Probability. 3rd ed. London: Oxford University Press, 1961

习　题　五

5.1　某单位指标在 10 月下半月完成一项露天作业，若半个月不下雨可按时完成任务，将获利 30 万元，若下雨则要亏 12 万元. 但若不开工(不论天气如何)均亏损 4 万元，试述此决策问题的状态集 $\boldsymbol{\Theta}$，行动集 Δ 和收益矩阵.

5.2　某厂购买一部机床，可使用三年，其一易损零件有备件供应，若随机床购买每只 250 元，以后再买每只 750 元. 在购买机床时应同时购买多少备件为宜的决策问题中其收益函数是什么？损失函数又是什么？若已知三年内最多需 4 个备用零件，写出其收益矩阵与损失矩阵.

5.3　某制药厂试制一种新的止痛片. 为了决定其投产量需了解此种新止痛片在止痛药市场中所占的比率 θ 是多少？这是个参数估计问题，若偏低估计 θ 将会导致供不应求，该赚到的钱没有赚到，造成工厂损失. 但偏高估计 θ 又导致供过于求，造成库存增加、资金积压，影响再生产. 这会给工厂造成更大损失，厂长认为供过于求引起的损失比供不应求引起的损失要高一倍，假如损失函数与 $|\theta-a|$ 成比例，请写出厂长所用的损失函数.

5.4　在二行动线性决策问题中的收益函数为

$$Q(\theta,a)=\begin{cases}b_1+m_1\theta, & a=a_1\\ b_2+m_2\theta, & a=a_2\end{cases}$$

在 $m_1>m_2$ 下，试写出相应的损失函数.

5.5 在缺少损失函数信息场合，可用参数 θ 处的密度函数值 $p(x|\theta)$ 与在行动 a 处的密度函数值 $p(x|a)$ 之间的距离来度量损失，如下两个距离较为常用.

(1) 熵的距离：$Le(\theta,a)=\mathrm{E}_{x|\theta}\left[\ln\dfrac{p(x|\theta)}{p(x|a)}\right]$；

(2) Hellinger 距离：$L_H(\theta,a)=\dfrac{1}{2}\mathrm{E}_{x|\theta}\left[\sqrt{\dfrac{p(x|a)}{p(x|\theta)}}-1\right]^2$.

假如 $X\sim N(\theta,1)$，证明：

$$Le(\theta,a)=\frac{1}{2}(a-\theta)^2$$

$$L_H(\theta,a)=1-\exp\{-(a-\theta)^2/8\}$$

5.6 设 $\chi=\{0,1,2\}$，在其上有一个二项分布族 $\{b(2,\theta),\theta$ 为 1/4 或 3/4$\}$. 若行动空间取为 $\Delta=\{1/4,3/4\}=\Theta$. 试写出所有的决策函数. 若取损失函数

$$L(\theta,a)=\begin{cases}0, & a=\theta\\ 1, & a\neq\theta\end{cases}$$

试计算每个决策函数的风险函数，并对它们作出比较.

5.7 设 $X_1,\cdots,X_n$ 是来自正态总体 $N(\mu,\sigma^2)$ 的一个样本，其中 σ^2 的估计有以下几种

$$\hat{\sigma}_1^2=\frac{1}{n-1}\sum(X_i-\overline{X})^2$$

$$\hat{\sigma}_2^2=\frac{1}{n+1}\sum(X_i-\overline{X})^2$$

$$\hat{\sigma}_3^2=\frac{1}{n+2}\sum X_i^2$$

试在损失函数 $L(\sigma^2,\hat{\sigma}_i^2)=\left(\dfrac{\hat{\sigma}^2}{\sigma^2}-1\right)$ 下分别计算它们的风险函数，并作出比较.

5.8 设 x 是来自二项分布 $b(n,\theta)$ 的一个观察值

$$\delta_1(x)=\frac{x}{n},\quad \delta_2(x)=\frac{x+\sqrt{n}/2}{n+\sqrt{n}}$$

是 θ 的两个估计. 试在平方损失下比较它们的风险函数.

5.9　设 $X_1,\cdots,X_n$ 是来自正态分布 $N(\mu,\sigma_0^2)$ 的一个样本,其中 σ_0^2 已知,在给定显著性水平 α 下,对假设检验问题

$$H_0: \mu=\mu_0 \quad 对 \quad H_1: \mu\neq\mu_0$$

作 u 检验. 其行动空间 $\Delta=\{0,1\}$, 0 表示接收, 1 表示拒绝. 试在损失函数

$$L(\mu,a)=\begin{cases}0, & 当\ \mu=\mu_0,a=0\ 或\ \mu\neq\mu_0,a=1\\ 1, & 其它\end{cases}$$

下计算 u 检验的风险函数.

5.10　在贝努里试验中记成功为 1, 失败为 0, 成功概率 θ 是未知的,但知有两种可能 $\theta_1=1/4$ 和 $\theta_2=1/2$.

(1) 通过一次贝努里试验结果 x 有下列四种决策函数

$$\delta_1(x)=\theta_1,\quad x=0,1$$
$$\delta_2(0)=\theta_1,\quad \delta_2(1)=\theta_2$$
$$\delta_3(0)=\theta_2,\quad \delta_3(1)=\theta_1$$
$$\delta_4(x)=\theta_2,\quad x=0,1$$

若取损失函数

$L(\theta,a)$	$a=1/4$	$a=1/2$
$\theta_1=1/4$	1	4
$\theta_2=1/2$	3	2

试寻找其中的最小最大决策函数;

(2) 通过两次贝努里试验结果之和 $x=x_1+x_2$ 可作出八个决策函数,试在相同损失函数下寻求其中的最小最大决策函数.

5.11　设随机变量 X 服从 Poisson 分布 $P(\lambda)$, 其中 λ 仅可取 $\lambda_1=1$ 和 $\lambda_2=2$ 中之一,现仅依据一次试验结果 x 构造如下两个决策函数作为 λ 的估计

$$\delta_1(x)=\begin{cases}1, & x=0,1\\ 2, & x=2,3,\cdots\end{cases}$$

$\delta_2(0)=1$，$\delta_2(1)$是随机变量，它以概率 2/3 取 1，而以概率 1/3 取 2；$\delta_2(2)$也是随机变量，它以概率 1/3 取 1，而以概率 2/3 取 2；最后 $\delta_2(x)=2, x=3,4,\cdots$.

试计算上述两个决策函数的风险函数.

5.12 一只罐里装有两个球，其中白球个数 θ 未知，通过返回抽样，可得一个容量为 2 的样本. 其中白球数记为 X. 据此定义如下的随机化决策函数 $\delta(x)$.

当 $X=0$ 时，有 $P(\delta(0)=0)=3/4, P(\delta(0)=1)=1/4$；

当 $X=1$ 时，$\delta(1)=1$；

当 $X=2$ 时，$P(\delta(2)=1)=1/4, P(\delta(2)=2)=3/4$.

试计算 $\delta(x)$的风险函数.

5.13 设 $X_1,\cdots,X_n$ 是来自二点分布 $b(1,\theta)$的一个样本，在平方损失下证明：样本均值 $\bar{x}$ 是 θ 的容许估计.

5.14 设 $X_1,\cdots,X_n$ 是来自 Poisson 分布 $P(\lambda)$的一个样本. 在平方损失下证明：样本均值 $\bar{X}$ 是 λ 的容许估计.

5.15 设 $X_1,\cdots,X_n$ 是来自正态分布 $N(\mu,1)$的一个样本，在平方损失函数下证明：样本均值 $\bar{x}$ 是 λ 的容许估计.

5.16 设 θ 是一批产品的不合格品率. 已知它必为 0.1 和 0.2 中之一，且其先验分布为

$$\pi(0.1)=0.7,\quad \pi(0.2)=0.3$$

假如从这批产品中随机取出 8 个进行检查，发现有 2 个是不合格品. 求 θ 的后验分布.

5.17 设 θ 是一批产品的不合格品率，从中任取 8 个产品进行检验，发现 3 个是不合格品. 要在下列先验分布的假设下，分别求 θ 的后验分布.

(1) $\theta\sim U(0,1)$；

(2) $\theta\sim\pi(\theta)=\begin{cases}2(1-\theta), & 0<\theta<1,\\ 0, & 其它.\end{cases}$

5.18 设 $X_1, X_2, \cdots, X_n$ 是来自密度函数 $p(x|\theta)$的一个样本. θ 的先验分布为 $\pi(\theta)$. 证明:按下述序贯方法亦可求出 θ 的后验分布:

1° 给定 X_1 的观察值 x_1 后,可求出 $\pi(\theta|x_1) \propto p(x_1|\theta)\pi(\theta)$;

2° 把 $\pi(\theta|x_1)$当作 θ 的先验分布,在给出 X_2 的观察值 x_2 后,可求出 $\pi(\theta|x_1, x_2) \propto p(x_2|\theta)\pi(\theta|x_1)$;

3° 按此方法重复,把上一步的后验分布作为下一步的先验分布,在给出观察值 $x_1, x_2, \cdots, x_{n-1}$后可得 $\pi(\theta|x_1, x_2, \cdots, x_{n-1})$最后可得

$$\pi(\theta|\underset{\sim}{x}) \propto p(x_n|\theta)\pi(\theta|x_1, x_2, \cdots, x_{n-1})$$

5.19 随机变量 X 服从均匀分布 $U\left(\theta-\frac{1}{2}, \theta+\frac{1}{2}\right)$,其中 θ 的先验分布为 $U(10,20)$. 假如获得 X 的一个观察值 12,求 θ 的后验分布. 假如连续获得 X 的 6 个观察值 11.0,11.5,11.7,11.1,11.4,10.9. 求 θ 的后验分布.

5.20 设随机变量 X 的密度函数为

$$p(x|\theta) = \frac{2x}{\theta^2}, \quad 0 < x < \theta < 1$$

(1) 假如 θ 的先验分布为 $U(0,1)$,求 θ 的后验分布;

(2) 假如 θ 的先验分布为

$$\pi(\theta) = 3\theta^2, \quad 0 < \theta < 1$$

求 θ 的后验分布.

5.21 验证 Poisson 分布的均值 λ 的共轭先验是 Gamma 分布.

5.22 验证正态方差 σ^2(均值已知)的共轭先验分布是倒 Gamma 分布.

5.23 设 $X=(X_1, \cdots, X_n)$是来自正态总体 $N(\mu, 1)$的一个样本. 假如 μ 的先验分布为 $N(0, \tau^2)$. 试分别用 X 的分布和充分统计量 $T = \sum_{i=1}^n X_i$ 的分布求 μ 的后验分布.

5.24 设 $X \sim N(\theta, \theta^2)$. 验证 θ 的共轭先验分布为倒正态分布,倒正态分布的密度函数为

$$\pi(\theta|\alpha, \mu, \tau) \propto |\theta|^{-\alpha}\exp\left\{-\left(\frac{1}{\theta} - \mu\right)^2 / 2\tau^2\right\}$$

5.25 多项分布 $M_k(n,\theta_1,\cdots,\theta_k)$中参数 $\theta=(\theta_1,\cdots,\theta_k)$的共轭先验分布是 Dirichlet 分布，其密度函数为

$$\pi(\theta_1,\cdots,\theta_k|\alpha_1,\cdots,\alpha_k)=\frac{\Gamma(\alpha_1+\cdots+\alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)}\prod_{i=1}^{k}\theta_i^{\alpha_i-1}I_s(\theta)$$

其中 I_s 为 $S=\{\theta=(\theta_1,\cdots,\theta_k):\sum_{i=1}^{k}\theta_i=1,\theta_i>0\}$的示性函数.

5.26 设某团体人的高度(单位:cm)服从均值为 θ,标准差为 5 的正态分布. 又设 θ 的先验分布为 $N(172.72,2.54)$，如今对随机选出的 10 人测量高度，其平均高度为 176.53 cm.

(1) 求 θ 的后验分布；

(2) 指出长为 2.5 cm 的区间，该区间内的点的后验密度值都大于区间外的点的后验密度值；

(3) 计算这个区间的概率.

5.27 从正态总体 $N(\theta,2^2)$中随机抽取容量为 100 的样本. 又设 θ 的先验分布为正态分布. 证明:不管先验标准差为多少，后验标准差一定小于 1.5.

5.28 设 $X_1,X_2,\cdots,X_n$ 是来自正态分布 $N(\theta,\sigma^2)$的一个样本. 令 $\theta_2=1/2\sigma^2$，又设(θ_1,θ_2)的联合分布如下给定：

1° 在固定 θ_2 时，θ_1 的条件分布为 $N(0,1/2\theta_2)$；

2° $\theta_2\sim\Gamma(\alpha,\lambda)$，

求(θ_1,θ_2)的后验分布 $\pi(\theta_1,\theta_2|x_1,x_2,\cdots,x_n)$.

5.29 某人每天早上在汽车站等公共汽车的时间(单位:分)服从均匀分布 $U(0,\theta)$，其中 θ 未知，假设 θ 的先验分布为

$$\pi(\theta)=\begin{cases}192/\theta^4, & \theta\geqslant 4\\ 0, & \theta<4\end{cases}$$

假如此人在三个早上等车的时间分别为 5,3,8 分钟，求 θ 的后验分布.

5.30 设 $X_1,X_2,\cdots,X_n$ 是来自均匀分布 $U(0,\theta)$的一个样本. 又设 θ 的先验分布为 Pareto 分布，其密度函数为

$$\pi(\theta)=\begin{cases}\alpha\theta_0^{\alpha}/\theta^{\alpha+1}, & \theta>\theta_0\\ 0, & \theta\leqslant\theta_0\end{cases}$$

其中参数 $\theta_0>0,\alpha>0$，证明:θ 的后验分布仍为 Pareto 分布，即 Pareto

分布是均匀分布 $U(0,\theta)$端点 θ 的共轭先验分布.

5.31　设参数 θ 的先验分布为 Beta 分布,已知其均值为 1/3,方差为 1/45,请确定先验分布.

5.32　设 λ 是指数分布 $Exp(\lambda)$ 中的参数,设 λ 的先验分布为 Gamma 分布,其均值为 0.0002,标准差为 0.0001,请确定先验分布.

5.33　设参数 θ 的先验分布为 Gamma 分布,若已知其上,下四分位数分别为 0.2 和 0.4,请确定其先验分布.

5.34　设随机变量 X 服从几何分布,即

$$P(X=k)=\theta(1-\theta)^k,\quad k=0,1,2,\cdots$$

其中参数 θ 的先验分布为均匀分布 $U(0,1)$.

(1) 若只对 X 作一次观察,观察值为 3,在平方损失函数下求 θ 的 Bayes 估计;

(2) 若对 X 作三次观察,观察值为 2,3,5,在平方损失函数下求 θ 的 Bayes 估计.

5.35　设为一顾客的服务时间(单位:分)服从指数分布 $Exp(\lambda)$,其中 λ 未知,又设 λ 的先验分布是均值为 0.2,标准差为 1 的 Gamma 分布.如今对 20 位顾客服务,平均服务时间是 3.8 分钟.在平方损失函数下,分别求 λ 和 $\theta=\lambda^{-1}$的 Bayes 估计.

5.36　设在 1200 英尺长的磁带上的缺陷数服从 Poisson 分布,其均值未知.又设 θ 的先验分布为 Gamma 分布 $Ga(3,1)$.对三盘磁带作检查,分别发现 2,0,6 个缺陷.在平方损失函数下,求 θ 的 Bayes 估计,并求其后验方差.

5.37　设 $X_1,X_2,\cdots,X_n$ 是来自 Poisson 分布 $P(\lambda)$的一个样本,又设 λ 的先验分布是均值为 μ_0 的 Gamma 分布.证明:λ 的后验均值是如下的加权平均

$$r_n \bar{X}_n+(1-r_n)\mu_0$$

并且当 $n\to\infty$时有 $r_n\to 1$.

5.38　设 θ 是不合格品率,它的先验分布为 Beta 分布 $Be(5,10)$.

(1) 假如随机检查 20 个产品,只发现 1 个不合格品.在平方损失函数下求 θ 的 Bayes 估计;

(2) 在检查 20 个产品中，不合格品为多少时才使后验风险最大？不合格品为多少时才使后验风险最小？这项讨论均在平方损失函数下进行.

5.39 设 $X_1, X_2, \cdots, X_n$ 是来自正态分布 $N(\theta, 2^2)$ 的一个样本，θ 的先验分布也是正态分布，其标准差为 1. 在平方损失函数下，最少要取多少样品才能使 θ 的 Bayes 估计的后验风险不超过 0.01?

5.40 设 $X_1, X_2, \cdots, X_n$ 是来自指数分布 $Exp(\lambda)$ 的一个样本，其中 λ 的先验分布为 $\pi(\lambda)$. 设 $\hat{\lambda}$ 是 λ 在平方损失函数 $L=(\hat{\lambda}-\lambda)^2$ 下的 Bayes 估计. 再命 $\theta=\lambda^2$，在平方损失函数 $L_1=(\hat{\theta}-\theta)^2$ 下，θ 的 Bayes 估计记为 $\hat{\theta}$，证明：$\hat{\theta}>\hat{\lambda}^2$.

5.41 设 θ 是一个城市中成年人赞成"公共场所禁止吸烟"的比例，对 θ 的先验分布两位统计学家有不同的建议：

$$\pi_A(\theta) = 2\theta, \qquad 0<\theta<1$$

$$\pi_B(\theta) = 4\theta^3, \qquad 0<\theta<1$$

随机抽样调查了该城市中 1000 名中年人，其中有 710 位投赞成票.

(1) 对 A 与 B 分别找出 θ 的后验分布；

(2) 对 A 与 B 分别找出 θ 的 Bayes 估计；

(3) 证明：在获得容量为 1000 的样本后，不管样本中投赞成票的人有多少，上述两个贝叶斯估计之差不会超过 0.002.

5.42 证明：在损失函数

$$L(\theta,\hat{\theta}) = \begin{cases} K_0(\theta-\hat{\theta}), & \text{当 } \hat{\theta}\leqslant\theta \\ K_1(\hat{\theta}-\theta), & \text{当 } \hat{\theta}>\theta \end{cases}$$

下 θ 的 Bayes 估计 $\hat{\theta}$ 是后验分布 $\pi(\theta|x)$ 的 $K_0/(K_0+K_1)$ 分位数.

5.43 设随机变量 $X\sim N(\theta,100)$，其中 θ 的先验分布为 $N(100, 225)$. 在损失函数

$$L(\theta,\hat{\theta}) = \begin{cases} 2(\theta-\hat{\theta}), & \text{当 } \hat{\theta}\leqslant\theta \\ (\hat{\theta}-\theta), & \text{当 } \hat{\theta}>\theta \end{cases}$$

下求 θ 的 Bayes 估计.

5.44 设 $X=(X_1,\cdots,X_n)$ 服从多项分布 $M(n,\theta_1,\cdots,\theta_k)$. 其中诸

$x_i \geqslant 0$, $\sum_{i=1}^{k} X_i = n, \theta_i \geqslant 0, \sum_{i=1}^{k} \theta_i = 1$. 现要求参数向量$(\theta_1, \cdots, \theta_k)$的 Jeffreys 先验分布.

5.45 设$\boldsymbol{X} = (X_1, \cdots, X_p)' \sim N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$,其$\boldsymbol{\Sigma}$为已知$p \times p$正定阵,又设$\boldsymbol{\theta}$的先验分布为$N_p(\boldsymbol{\mu}, \boldsymbol{A})$,其中$\boldsymbol{\mu}$为已知$p$维向量,$\boldsymbol{A}$为已知$p \times p$的正定阵.

(1) 证明:θ的后验分布仍为p元正态,其均值向量为$\boldsymbol{x} - \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \boldsymbol{A})^{-1}(\boldsymbol{x} - \mu)$,方差与协方差阵为$(\boldsymbol{\Sigma}^{-1} + \boldsymbol{A}^{-1})^{-1} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}(\boldsymbol{A} + \boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}$;

(2) 在正定二次损失函数

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}) = (\boldsymbol{\delta} - \boldsymbol{\theta})'\boldsymbol{Q}(\boldsymbol{\delta} - \boldsymbol{\theta})$$

下寻求$\boldsymbol{\theta}$的 Bayes 估计.

5.46 设某设备的寿命X服从指数分布,其密度为

$$p(x|\lambda) = \lambda e^{-\lambda x}, \quad x > 0, \lambda > 0$$

对同类设备作n次观察,得$x_1, \cdots, x_n$,如今规定该设备的可靠性为$g(t_0, \theta) = e^{-t_0\theta}$,$t_0$是已知正实数.若设$\theta$的先验分布为 Gamma 分布$Ga(\alpha, 1/\tau)$,试在平方损失函数下求$g(t_0, \theta)$的 Bayes 估计.

5.47 设$X \sim N(\theta, \sigma^2)$,其中$\sigma$已知,若已知$\theta$为正,其先验分布为无信息先验分布$\pi(\theta) = I_{(0,\infty)}(\theta)$.试求其后验密度函数.并在平方损失函数下求$\theta$的 Bayes 估计.

5.48 设有一批产品,其不合格品率为p.若将每N件装为一箱.现从一箱中随机抽检n件产品,得知不合格品数是r.试求这箱的不合格品率$q = w/N$的 Bayes 估计.其中w是这箱中的不合格品数,假如取平方损失函数$L(w, \delta) = (\delta - w/N)^2$.

5.49 设$X \sim b(n, p)$,$0 < p < 1$,证明:$\delta = x/n$在损失函数$L(\theta, \delta) = (\delta - \theta)^2/\theta(1-\theta)$下是$\theta$的最小最大估计(提示:先验分布取均匀分布).

5.50 设$X \sim N(\theta, 1)$,$\theta \in \mathbf{R}$,证明:在一次试验下,$\delta(x) = x$是在平方损失函数下θ的最大最小估计.(提示:$\theta \sim N(0, \tau^2)$)

5.51 对正态分布$N(\theta, 1)$观察,获得三个独立观察值

$$X_1 = 2, X_2 = 3, X_3 = 4$$

若θ的先验分布为$N(3, 1)$,求θ的0.95可信区间.

5.52 设 $X_1,X_2,\cdots,X_n$ 是来自 Poisson 分布 $P(\lambda)$ 的一个样本，假如 λ 的先验分布为 Gamma 分布 $Ga(a,b)$，求 λ 的 $1-\alpha$ 可信下限.

5.53 设 $X_1,X_2,\cdots,X_n$ 是来自均匀分布 $U(0,\theta)$ 的一个样本，其中 θ 的先验分布为 Pareto 分布，其密度函数为

$$\pi(\theta)=\frac{\alpha\theta_0^\alpha}{\theta^{\alpha+1}},\quad \theta>\theta_0$$

其中 $\theta_0>0,\alpha>0$ 是两个已知常数.

(1) 在平方损失函数下求 θ 的 Bayes 估计；

(2) 求 θ 的 $1-\alpha$ 可信上限.

5.54 设 $X_1,\cdots,X_n$ 是来自如下 Cauchy 分布 $C(\theta,1)$ 的一个样本

$$p(x|\theta)=[1+(\theta-x)^2]^{-1},\quad x\in\mathbf{R},\theta>0$$

假如 θ 取无信息先验分布 $\pi(\theta)=1,\theta>0$. 寻求 θ 的后验分布，假如从 $C(\theta,1)$ 获得 5 个观察值：4.0，5.5，7.3，4.5，3.0. 试给出 θ 的 95%HPD 可信区间.

第六章　统计计算方法

在统计领域里，统计计算技术近年来发展很快，它使得许多统计方法，尤其是 Bayes 统计得到广泛运用. 下面我们介绍一些基本的计算方法以及最近发展很快的一些计算技术.

§6.1　随机数的产生

在各种统计计算中常需要产生各种概率分布的随机数，而大多数概率分布的随机数的产生均基于均匀分布 $U(0,1)$ 的随机数. 本节我们首先介绍产生各种分布随机数的方法. 基本方法有如下三种：逆变换方法，合成方法和筛选方法. 然后介绍一些特定分布的随机数产生方法. 此处，我们假定读者已掌握了 $U(0,1)$ 分布的随机数的产生方法. 事实上，$U(0,1)$ 分布随机数的产生在大多数计算语言（软件）中都有现成的程序可调用，另外，$U(0,1)$ 分布随机数的产生在一些文献[1][2]中有详细的介绍.

本节主要介绍一个随机数的产生方法. 一般情况下，n 个独立随机变量可由 n 次重复抽样获得.

§6.1.1　逆变换法

设随机变量 X 的分布函数为 $F(x)$，定义

$$F^{-1}(y) = \inf\{x : F(x) \geqslant y\}, \quad 0 \leqslant y \leqslant 1 \tag{6.1}$$

则有如下定理.

定理 6.1　设随机变量 U 服从 $(0,1)$ 上的均匀分布，则 $X = F^{-1}(U)$ 的分布函数为 $F(x)$.

证明： 由(6.1)式和均匀分布的分布函数可得

$$P(X \leqslant x) = P(F^{-1}(U) \leqslant x)$$
$$= P(U \leqslant F(x)) = F(x)$$

证毕.

由定理 6.1,要产生来自 $F(x)$ 的随机数,只要先产生来自 $U(0,1)$ 的随机数 u,然后计算 $F^{-1}(u)$ 即可. 其具体步骤为

$$\begin{cases}(1)\ 由\ U(0,1)\ 抽取\ u, \\ (2)\ 计算\ x = F^{-1}(u),其中\ F^{-1}\ 如(6.1)中定义.\end{cases} \tag{6.2}$$

例 6.1 设 $X \sim U(a,b)$,则其分布函数

$$F(x) = \begin{cases} 0, & x < a \\ \dfrac{x-a}{b-a}, & a \leqslant x < b \\ 1, & x \geqslant b \end{cases}$$

从而

$$F^{-1}(y) = a + (b-a)y, \quad 0 \leqslant y \leqslant 1$$

若由 $U(0,1)$ 抽得 u,则 $a+(b-a)u$ 是来自 $U(a,b)$ 的一个随机数.

例 6.2 设 $X_1,\cdots,X_n$ 是来自 $F(x)$ 的一个样本,$Y_1=\min(X_1,\cdots,X_n)$,我们要产生 Y_1 的随机数. 当然,这可以通过产生 n 个来自 $F(x)$ 的独立变量并求最小值来得到,但这样做是不必要的. 我们知道,

$$F_{Y_1}(y) = 1 - [1 - F(y)]^n \triangleq G(y)$$

易计算

$$Y_1 = G^{-1}(U) = F^{-1}(1 - (1 - U)^{1/n})$$

应用(6.2),可直接产生来自 $G(y)$ 的随机数. 譬如,若 $F(x)=x$, $0 \leqslant x \leqslant 1$,则 $Y_1 = 1-(1-U)^{1/n}$. 由于 $1-U$ 与 U 同分布,上式也可写成 $Y_1 = 1-U^{1/n}$.

例 6.3 设连续随机变量 X 的密度函数为

$$p(x) = \begin{cases} c_i, & x_i \leqslant x < x_{i+1}, \quad i = 0,\cdots,n-1 \\ 0, & 其它 \end{cases}$$

其中 $c_i > 0$, $a = x_0 < \cdots < x_n = b$($a,b$ 可为无穷), $\int_a^b p(x)\mathrm{d}x = 1$.

令 $p_i = \int_a^{x_i} f(x)\mathrm{d}x$, $i=1,\cdots,n$,则对任一 x,令 $i=\max\{j: x_j \leqslant x\}$,有

$$F(x) = p_i + c_i(x - x_i)$$

由 $F(X)=U$ 可解出

$$X = x_i - (U - p_i)/c_i$$

此处 i 满足 $p_i \leqslant U < p_{i+1}$. 具体步骤为

$$\begin{cases}(1)\ 产生\ U(0,1)\ 随机数\ u, \\ (2)\ 确定\ i,使\ p_i \leqslant u < p_{i+1}, \\ (3)\ 计算\ x = x_i + (u - x_i)/c_i.\end{cases}$$

§6.1.2 合成法

合成方法的应用最早见于 Butler 的书[3]中. 他的想法是:如果 X 的密度函数 $p(x)$ 难于抽样,而 X 关于 Y 的条件密度函数 $p(x|y)$ 以及 Y 的密度函数 $g(y)$ 均易于抽样,则 X 的随机数可如下产生.

$$\begin{cases}(1)\ 由\ Y\ 的分布\ g(y)\ 抽取\ y, \\ (2)\ 由条件分布\ p(x|y)\ 抽取\ x.\end{cases} \tag{6.3}$$

可以证明[2],由(6.3)得到的 X 服从 $p(x)$.

例 6.4 设 X 的密度函数为 $p(x) = \sum_{i=1}^n \alpha_i p_i(x)$,其中诸 $\alpha_i > 0$, $\sum_{i=1}^n \alpha_i = 1$, $p_i(x)$ 是密度函数. 令 $\alpha_0 = 0$,由合成法,X 的随机数可如下抽取:

$$\begin{cases}(1)\ 由\ U(0,1)\ 抽取\ u, \\ (2)\ 确定\ i,使 \sum\limits_{j=0}^{i-1}\alpha_j < u \leqslant \sum\limits_{j=0}^{i}\alpha_j, \\ (3)\ 由\ p_i(x)\ 抽取\ x.\end{cases}$$

譬如,$p(x) = (1+2x)/6$, $0<x<2$,其分布函数为 $F(x) = (x+x^2)/6$,若用逆变换法抽样,则要解二次方程,较为麻烦. 考虑用合成法,将 $p(x)$ 分解,$p_1(x) = 1/2$, $0<x<2$, $p_2(x) = x/2$, $0<x<2$, $\alpha_1 = 1/3$, $\alpha_2 = 2/3$,由(6.3),结合逆变换法,我们可给出具体抽样步骤为

$$\begin{cases}(1)\ \text{由 } U(0,1) \text{ 独立地抽取 } u_1, u_2, \\ (2)\ \text{计算 } x = \begin{cases} 2u_2, & u_1 < 1/3, \\ 2\sqrt{u_2}, & u_1 \geqslant 1/3. \end{cases}\end{cases}$$

§6.1.3 筛选抽样

假设我们要从 $p(x)$抽样，如果可以将 $p(x)$表示成 $p(x)=c\cdot h(x)\cdot g(x)$，其中 $h(\cdot)$是一个密度函数且易于抽样，而 $0<g(x)\leqslant 1$，$c\geqslant 1$是常数，则 X 的抽样可如下进行.

$$\begin{cases}(1)\ \text{由 } U(0,1) \text{ 抽取 } u, \text{由 } h(y) \text{ 抽取 } y, \\ (2)\ \text{如果 } u \leqslant g(y), \text{则 } x = y, \text{停止}, \\ (3)\ \text{如果 } u > g(y), \text{回到(1)}. \end{cases} \tag{6.4}$$

由(6.4)表示的抽样方法称为筛选抽样(Acceptance/Rejection Sampler)，其理论依据是下述定理 6.2

定理 6.2 设 X 的密度函数为 $p(x)$，且 $p(x)=c\cdot h(x)\cdot g(x)$，其中 $c\geqslant 1, 0<g(x)\leqslant 1, h(\cdot)$是一个密度函数. 令 U 和 Y 分别服从 $U(0,1)$和 $h(y)$，则在 $U\leqslant g(Y)$的条件下，Y 的条件密度为

$$p_Y(x|U \leqslant g(Y)) = p(x)$$

证明：由 Bayes 公式，注意到 $1/c = \int g(x)h(x)\mathrm{d}x$，有

$$\begin{aligned} p_Y(x|U \leqslant g(Y)) &= \frac{P(U \leqslant g(Y)\,|\,Y = x)h(x)}{P(U \leqslant g(Y))} \\ &= \frac{P(U \leqslant g(x))h(x)}{\int P(U \leqslant g(Y)\,|\,Y = x)h(x)\mathrm{d}x} \\ &= \frac{g(x)h(x)}{\int g(x)h(x)\mathrm{d}x} = p(x) \end{aligned}$$

证毕.

应该指出，由(6.4)抽样有两点很重要. 一方面 $h(x)$应易于抽样，另一方面，我们不能保证抽几次可以得到一个 $p(x)$的随机数，这就涉

及到关于抽样方法的效的问题. 所谓效，指的是平均抽几次可以得到一个随机数[24].

$h(x)$的选取有多种方法. 一种直观的方法是：如果存在一个函数$M(x)$，满足$p(x)\leqslant M(x)$，且$c=\int M(x)\mathrm{d}x<\infty$，令$h(x)=M(x)/c$，则$f(x)=c\cdot h(x)\cdot p(x)/M(x)$. 若$h(x)$易于抽样，则(6.4)变成

$$\begin{cases}(1)\ \text{由 } U(0,1) \text{ 抽取 } u, \text{由 } h(y) \text{ 抽取 } y,\\ (2)\ \text{如果 } u \leqslant p(y)/M(y), \text{则 } x=y, \text{停止},\\ (3)\ \text{如果 } u > p(y)/M(y), \text{回到(1)}.\end{cases} \tag{6.5}$$

特别，若X的取值空间有限，譬如$X\sim p(x)$，$-\infty<a\leqslant x\leqslant b<\infty$，并设$M=\sup p(x)$存在，则可取$h(x)=1/(b-a)$，$c=M(b-a)$，$g(x)=p(x)/M$，(6.5)简化为

$$\begin{cases}(1)\ \text{由 } U(0,1) \text{ 独立抽取 } u_1, u_2,\\ (2)\ \text{计算 } y=a+u_2(b-a),\\ (3)\ \text{如果 } u_1 \leqslant g(y)=p(a+u_2(b-a))/M, \text{则 } x=y, \text{停止},\\ (4)\ \text{如果 } u_1 > g(y), \text{回到(1)}.\end{cases} \tag{6.6}$$

例 6.5　设$p(x)=4x^3$，$0\leqslant x\leqslant 1$. 此处$M=4$，$a=0$，$b=1$，$h(x)=1$，$g(x)=x^3$. 应用(6.6)即可给出如下抽样方法：

$$\begin{cases}(1)\ \text{由 } U(0,1) \text{ 独立抽取 } u_1, u_2,\\ (2)\ \text{若 } u_1 \leqslant u_2^3, \text{则 } x=u_2, \text{停止},\\ (3)\ \text{若 } u_1 > u_2^3, \text{转到(1)}.\end{cases}$$

例 6.6　设

$$X\sim p(x)=\frac{1}{\Gamma(\alpha)}x^{\alpha-1}\mathrm{e}^{-x},\quad x>0 \tag{6.7}$$

$0<\alpha<1$ 已知. 注意到

$$\frac{1}{\Gamma(\alpha)}x^{\alpha-1}\mathrm{e}^{-x}\leqslant M(x)=\begin{cases}x^{\alpha-1}/\Gamma(\alpha), & 0<x\leqslant 1\\ \mathrm{e}^{-x}/\Gamma(\alpha), & x>1\end{cases}$$

因

$$c = \int_0^{\infty} M(x)\mathrm{d}x = (1/\alpha + 1/\mathrm{e})/\Gamma(\alpha) < \infty$$

取

$$h(x) = \begin{cases} x^{\alpha-1}/(1/\alpha + 1/\mathrm{e}), & 0 < x \leqslant 1 \\ \mathrm{e}^{-x}/(1/\alpha + 1/\mathrm{e}), & x > 1 \end{cases}$$

则

$$g(x) = \begin{cases} \mathrm{e}^{-x}, & 0 < x \leqslant 1 \\ x^{\alpha-1}, & x > 1 \end{cases}$$

于是，分布(6.7)的随机数可如下抽取：

$$\begin{cases} (1)\ 由\ U(0,1)\ 抽取\ u, \\ (2)\ 由\ h(y)\ 抽取\ y(可使用逆变换方法,见习题\ 6.2), \\ (3)\ 当\ y \in (0,1]\ 时,如果\ u \leqslant \mathrm{e}^{-y}\ 则\ x = y,否则转到(1), \\ (4)\ 当\ y > 1\ 时,如果\ u < y^{\alpha-1}\ 则\ x = y,否则转到(1). \end{cases}$$

筛选抽样方法是一个相当重要的抽样方法，它可解决许多难以直接抽样的分布的抽样问题. 关于筛选抽样的进一步讨论查阅有关文献[4].

§ 6.1.4 连续分布的抽样方法

§ 6.1.4.1 指数分布的抽样法

设 $X \sim p(x) = (1/\theta)\mathrm{e}^{-x/\theta}I_{(0<x)}$， $\theta>0$ 已知，则 $X/\theta \sim Exp(1)$，因此，我们只需讨论标准指数分布 $Exp(1)$的抽样方法.

对 $X \sim p(x) = \mathrm{e}^{-x}I_{(x>0)}$，由逆变换法有

$$U = F(X) = 1 - \mathrm{e}^{-X}$$

即

$$X = -\ln(1-U)$$

由于 $1-U$ 与 U 同分布，上式也可以写成 $X = -\ln U$，于是给出如下抽样方法：

$$\begin{cases} (1)\ 由\ U(0,1)\ 抽取\ u, \\ (2)\ 计算\ x = -\ln u. \end{cases} \tag{6.8}$$

由(6.8)式抽样看上去简单,但因在计算机上计算自然对数需用到级数展开,每得一个指数分布随机数都要用到一次级数展开,是较费时的,因而效率不高.

我们知道,标准指数变量与 $U(0,1)$ 变量有紧密的联系,我们可用之于同时产生多个独立 $Exp(1)$ 变量.这用到下述引理.

引理 6.3　设 $X_1,\cdots,X_n$,　$Y_1,\cdots,Y_{n-1}$ 是来自 $U(0,1)$ 的随机样本,$Y_{(1)}<\cdots<Y_{(n-1)}$ 是 $Y_1,\cdots,Y_{n-1}$ 的次序统计量.记 $Y_{(0)}=0,Y_{(n)}=1$,令 $Z_k=(Y_{(k-1)}-Y_{(k)})\ln(\prod_{i=1}^n X_i)$,　$k=1,\cdots,n$,则 $Z_1,\cdots,Z_n$ 独立同分布,服从 $Exp(1)$.

证明:因为 $W=-2\ln(\prod_{i=1}^n X_i)\sim\chi^2(2n)$.由 $(Y_{(1)},\cdots,Y_{(n-1)})$ 与 W 的独立性和次序统计量的分布得 $(Y_{(1)},\cdots,Y_{(n-1)},W)\propto(n-1)w^{n-1}e^{-w/2}$,$0<y_{(1)}\leqslant\cdots\leqslant y_{(n-1)}<1$.由给出的 $(Z_1,\cdots,Z_n)$ 到 $(Y_{(1)},\cdots,Y_{(n-1)},W)$ 变换可得

$$W=2\sum_{i=1}^n Z_i,\quad \frac{\partial(Z_1,\cdots,Z_n)}{\partial(y_{(1)},\cdots,y_{(n-1)},w)}=\frac{w^{n-1}}{2^n}$$

所以可得 $(Z_1,\cdots,Z_n)\propto\exp\left\{-\sum_{i=1}^n Z_i\right\}$,证毕.

由引理 6.3,我们可给出下述抽样法:

$$\begin{cases}(1)\ 由\ U(0,1),独立抽取\ 2n-1\ 个随机数\ u_1,\cdots,u_{2n-1},\\(2)\ 计算\ r=\ln(\prod_{i=1}^n u_i),\\(3)\ 将\ u_{n+1},\cdots,u_{2n-1}\ 按从小到大排序,记为\ y_{(1)}<\cdots<y_{(n-1)},\\(4)\ 计算\ x_k=(y_{(k-1)}-y_{(k)})r,其中\ y_{(0)}=0,y_{(n)}=1.\end{cases}\tag{6.9}$$

则 $x_1,\cdots,x_n$ 是来自 $Exp(1)$ 的随机数.

我们可将(6.9)与(6.8)进行比较.若我们要抽 n 个 $Exp(1)$ 变量,用(6.8)只要抽 n 个 $U(0,1)$ 变量,但要进行 n 次自然对数的计算,用(6.9)虽然要抽 $2n-1$ 个 $U(0,1)$ 变量,但只要进行一次自然对数的计算,然而,用(6.9)还要进行一次 $n-1$ 个数据的排序.在计算机上的实

际运行表明，用(6.9)一般要比(6.8)快，见习题 6.7.

指数分布的抽样还有许多其它方法.有文献中利用筛选法等给出 $Exp(1)$的另两种抽样方法[4]，此处从略.

§6.1.4.2 正态分布的抽样

基于与指数分布一样的原因，我们只需要讨论标准正态分布 $N(0,1)$的抽样问题.

Box 和 Muller 在 1958 年给出一种用变换产生 $N(0,1)$随机数的方法[5]，其理论基础是下面的引理.

引理 6.4 设 U_1,U_2 是独立同分布的 $U(0,1)$变量，令

$$X_1 = (-2\ln U_1)^{1/2}\cos(2\pi U_2) \tag{6.10}$$

$$X_2 = (-2\ln U_1)^{1/2}\sin(2\pi U_2) \tag{6.11}$$

则 X_1 与 X_2 独立，均服从标准正态分布.

证明：见习题 6.8.

引理 6.4 给出同时产生两个 $N(0,1)$随机数的方法：

$$\begin{cases} (1)\ 由\ U(0,1)\ 独立抽取\ u_1,u_2, \\ (2)\ 用(6.10),(6.11)\ 计算\ x_1,x_2. \end{cases} \tag{6.12}$$

$N(0,1)$变量的产生有许多方法.比如，我们可结合合成法和筛选法抽取 $N(0,1)$变量如下：首先将 X 固定在 $X\geqslant 0$ 上，此时 X 的(条件)分布密度函数为

$$p(x) = \sqrt{\frac{2}{\pi}}e^{-x^2/2}, \quad x \geqslant 0 \tag{6.13}$$

将(6.13)改写为 $p(x)=c\cdot h(x)\cdot g(x)$，其中，

$$c = \sqrt{2e/\pi}, \quad h(x) = e^{-x}, \quad g(x) = e^{-(x-1)^2/2} \tag{6.14}$$

应用筛选法可得到来自(6.13)的随机数，其中筛选条件为 $U\leqslant g(Y)=e^{-(Y-1)^2/2}$，其等价条件为 $-\ln U\geqslant (Y-1)^2/2$，而 $-\ln U$ 是服从标准指数分布的.又因为对 $N(0,1)$变量而言，$X>0$ 与 $X<0$ 是同等可能的，于是用合成法可得到 $N(0,1)$变量.具体步骤如下：

$$\begin{cases} (1)\ \text{由}\ Exp(1)\ \text{独立抽取}\ v_1, v_2, \\ (2)\ \text{若}\ v_1 \geqslant (v_2 - 1)^2/2, \text{进入(3)，否则转到(1)}, \\ (3)\ \text{从}\ U(0,1)\ \text{抽}\ u, \\ (4)\ x = \begin{cases} v_2, & \text{如果}\ u \leqslant 0.5, \\ -v_2, & \text{如果}\ u > 0.5. \end{cases} \end{cases} \tag{6.15}$$

$N(0,1)$变量的抽取还有一个著名的近似方法，即用 n 个 $U(0,1)$ 变量产生一个 $N(0,1)$变量，其理论基础是中心极限定理，抽样公式为 $x=\sqrt{12n}(\bar{u}-1/2)$，其中 $\bar{u}=\frac{1}{n}\sum_{i=1}^{n}u_i$，$u_1,\cdots,u_n$ 是抽自 $U(0,1)$的随机数. 在实用中常取 $n=6$ 或 $n=12$. 关于 $N(0,1)$的抽样方法的进一步的介绍可参阅有关文献[2],[4].

§6.1.4.3 Gamma 分布的抽样方法

Gamma 分布(以及倒 Gamma 分布)的抽样在统计计算中扮演了一类重要角色. 注意到，若 $X\sim Ga(\alpha,\lambda)$，则 $X/\lambda\sim Ga(\alpha,1)$，因此只需讨论 $Ga(\alpha,1)$的抽样方法.

若 $\alpha=1$，$Ga(1,1)$即 $Exp(1)$，它的抽样方法我们已经介绍过. 若 α 是整数 m，由可加性不难给出 $Ga(m,1)$变量的抽样方法：

$$\begin{cases} (1)\ \text{独立地抽}\ m\ \text{个}\ Exp(1)\ \text{变量}\ r_1,\cdots,r_m, \\ (2)\ \text{计算}\ x = \sum_{i=1}^{m} r_i. \end{cases} \tag{6.16}$$

如果 $0<\alpha<1$，我们在例 6.6 中已经用筛选法给出了 $Ga(\alpha,1)$的抽样方法，此处主要介绍 $\alpha>1$ 且 α 不是整数时 $Ga(\alpha,1)$的抽样方法.

显然，$\alpha>1$ 时的 $Ga(\alpha,1)$变量可以转化为多个 $Ga(\beta,1)$，$0<\beta<1$ 变量的和，就是说，找一个整数 m(比如 $m=[\alpha]+1$)，令 $\beta=\alpha/m$，则 $0<\beta<1$，于是我们可以产生 m 个 $Ga(\beta,1)$变量，将之求和即可得到一个 $Ga(\alpha,1)$变量. 这种方法是笨重的，尤其是 α 比较大时，可以想象它的效率很低.

Wallace 在 1974 年构造了一个筛选方法产生 $Ga(\alpha,1)$变量，其方法如下：设 $p(x)=x^{\alpha-1}e^{-x}/\Gamma(\alpha)$，记 $m=[\alpha]$，$p=\alpha-m$，$0<p<1$，令

$$h(x) = p\frac{x^{m-1}e^{-x}}{(m-1)!} + (1-p)\frac{x^m e^{-x}}{m!}, \quad x > 0$$

$$c = \frac{(m-1)!\,m^p}{\Gamma(\alpha)}$$

$$g(x) = \left(\frac{x}{m}\right)^p \left[1 + (1-p)\left(\frac{x}{m} - 1\right)\right]^{-1}$$

不难验证有 $p(x)=c\cdot h(x)\cdot g(x)$，由此给出 $Ga(\alpha,1)$ 的抽样方法：

(1) 由 $U(0,1)$ 抽取 r，

(2) 若 $r \leqslant p$，抽 m 个 $U(0,1)$ 变量 $u_1,\cdots,u_m$，

  计算 $v = -\ln(\prod_{i=1}^{m} u_i)$，转到(4)，

(3) 若 $r > p$，抽 $m-1$ 个 $U(0,1)$ 变量 $u_1,\cdots,u_{m+1}$， (6.17)

  计算 $v = -\ln(\prod_{i=1}^{m+1} u_i)$，

(4) 由 $U(0,1)$ 抽 u，若 $u \leqslant (v/m)^p[1 + (v/m - 1)(1-p)]^{-1}$，

  则 $x = v$，停止，

(5) 若 $u > (v/m)^p[1 + (v/m - 1)(1-p)]^{-1}$，转到(1).

§ 6.1.4.4 Beta 分布的抽样方法

由熟知的 Beta 分布与 Gamma 分布的关系：若 $X_1 \sim Ga(\alpha,1)$，$X_2 \sim Ga(\beta,1)$，且 X_1 与 X_2 独立，则 $X = X_1/(X_1+X_2) \sim Be(\alpha,\beta)$，由此可给出 Beta 分布的一个抽样方法：

(1) 由 $Ga(\alpha,1)$ 抽 x_1，由 $Ga(\beta,1)$ 抽 x_2，

(2) 计算 $x = x_1/(x_1 + x_2)$. (6.18)

如果 α,β 是整数，设 $U_1,\cdots,U_{\alpha+\beta-1}$ 是来自 $U(0,1)$ 的随机变量，则其第 α 个次序统计量 $U_{(\alpha)}$ 服从 $Be(\alpha,\beta)$ 分布，从而可得到它的随机数：

(1) 产生 $\alpha+\beta-1$ 个 $U(0,1)$ 随机数 $u_1,\cdots,u_{\alpha+\beta-1}$，

(2) 将 $u_1,\cdots,u_{\alpha+\beta-1}$ 从小到大排序，记为 $u_{(1)},\cdots,u_{(\alpha+\beta-1)}$，

(3) $x = u_{(\alpha)}$.

对一般的 α,β，最简单的抽样方法或许是筛选抽样. 因 $p(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$，由求导不难得出当 $x_0 = (\alpha-1)/(\alpha+\beta-2)$ 时

$p(x)$达到最大. 记

$$M = p(x_0) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\frac{\alpha-1}{\alpha+\beta-2}\right)^{\alpha-1}\left(\frac{\beta-1}{\alpha+\beta-2}\right)^{\beta-1} \tag{6.19}$$

类似于(6.6)式,Beta 分布的抽样步骤如下:

$$\begin{cases}(1)\ 由\ U(0,1)\ 分布独立抽取\ u_1,u_2,\\(2)\ 如果\ u_1 \leqslant \dfrac{1}{M}\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}u_2^{\alpha-1}(1-u_2)^{\beta-1},则\ x=u_2,停止,\\(3)\ 否则,回到(1),其中\ M\ 由(6.19)确定.\end{cases} \tag{6.20}$$

若$(\alpha-1)/(\alpha+\beta-2)<0$,最大点在 0 或 1 达到,相应改变$M$,上面的方法同样可行. 当无最大点时(如 $\alpha=1,\beta<1$ 或 $\alpha<1,\beta=1$),按幂分布产生随机数(参考例 6.5).

§6.1.4.5　其它连续分布的抽样

一般的抽样方法已在 A,B,C 中加以介绍,对数正态分布,χ^2 分布,t 分布以及 F 分布的抽样均可由正态分布的抽样导出,此处不再叙述.

§6.1.5　离散分布的抽样方法

设 X 为一离散变量,其概率函数为

$$P(X=k)=p_k,\quad k=0,1,\cdots$$

记 $g_k=\sum_{i=0}^{k}p_i,\quad k=0,1,\cdots$,又设$U$为$U(0,1)$变量,则

$$P(g_{k-1}<U\leqslant g_k)=\int_{g_{k-1}}^{g_k}\mathrm{d}u=g_k-g_{k-1}=p_k,\quad k=0,1,2,\cdots$$

其中 $g_{-1}=0$. 于是由逆变换法,$X=\min\{k:g_{k-1}<U\leqslant g_k\}$,具体抽样步骤为

$$\begin{cases}(1)\ 由\ U(0,1)\ 抽取\ u,置初值\ k=0,\\(2)\ 若\ u\leqslant g_k,则\ x=k,停止,\\(3)\ 若\ u>g_k,令\ k=k+1,转到(2).\end{cases} \tag{6.21}$$

容易计算算法(6.21)的平均运行次数为

$$C = 1 + \sum_{k=0}^{\infty} k p_k = 1 + \mathrm{E}(X) \tag{6.22}$$

在许多场合，我们可以通过改变对 X 的搜索起点值而改进算法(6.21)的效，亦即减少平均运行次数. 比如我们可由 $X=m$(m 为众数，中位数等)开始搜索，其步骤如下

$$\begin{cases} (1)\ \text{由 } U(0,1) \text{ 抽取 } u, \text{若 } u > g_m, \text{转}(4), \\ (2)\ k = m-1, \text{若 } u > g_k, \text{则 } x = k, \text{停止，否则} \\ (3)\ k = k-1, \text{回到}(2), \\ (4)\ k = m+1, \text{若 } u \leqslant g_k, \text{则 } x = k, \text{停止，否则} \\ (5)\ k = k+1, \text{转到}(4). \end{cases} \tag{6.23}$$

不难看出，若抽样结果为 k，则运行次数为

$$\begin{cases} 2 + (m-k), \text{ 如果 } k = 0,1,\cdots m \\ 1 + (k-m), \text{ 如果 } k = m+1, m+2, \cdots \end{cases}$$

于是平均运行次数为

$$\begin{aligned} C &= \sum_{k=0}^{m} [2 + (m-k)] p_k + \sum_{k=m+1}^{\infty} [1 + (k-m)] p_k \\ &= (m+1)\sum_{k=0}^{m} p_k + 1 - \sum_{k=0}^{m} k p_k + \sum_{k=m-1}^{\infty} k p_k - m \sum_{k=m+1}^{\infty} p_k \\ &= (m+1) g_m + 1 - m(1-g_m) - 2\sum_{k=0}^{m} k p_k + \mathrm{E}(X) \\ &= 1 + \mathrm{E}(X) - \gamma(m) \end{aligned}$$

其中 $\gamma(m) = 2\sum_{k=0}^{m} k p_k + m - (2m+1) g_m$. 如果 $\gamma(m) > 0$，则算法(6.23)比(6.21)有效. 下面的定理说明什么情况下有 $\gamma(m)>0$.

引理 6.5　在上述场合下，若存在 $m>0$，满足

$$p_0 < \sum_{k=1}^{m} [(2k-1) p_k], \quad g_m \leqslant 1/2$$

则 $\gamma(m) > 0$.

证明：由条件 $p_0 < \sum_{k=1}^{m} (2k-1) p_k$，有

$$2\sum_{k=0}^{m} k p_k - g_m > 0$$

又 $g_m \leqslant 1/2$，即 $m-2mg_m \geqslant 0$，由此即有 $\gamma(m)>0$.

注：若 $p_0 < \sum_{k=1}^{m} p_k$，则有 $p_0 < \sum_{k=1}^{m}(2k-1)p_k$ 成立.

引理 6.6　记 $m_0 \doteq \max\{m: g_m \leqslant 1/2\}$，如果 $g_{m_0}+g_{m_0+1} \leqslant 1$，则 $\gamma(m)$ 在 m_0 处达到最大；如果 $g_{m_0}+g_{m_0+1}>1$，则 $\gamma(m)$ 在 m_0+1 处达到最大.

证明：由 $\gamma(m+1)-\gamma(m)=1-g_m-g_{m+1}$ 立见.

例 6.7　设 $X \sim b(n,p)$，当 n 不是很大时，常用的抽样方法有如下几种：

(a)变换方法. 由二项分布的可加性，X 可以表示为 n 个二点分布变量的独立和，因此，我们可以先产生 n 个 $b(1,p)$ 分布的随机数，然后相加即可得到一个 $b(n,p)$ 的随机数. 具体步骤为

$$\begin{cases} (1)\ \text{产生 } n \text{ 个 } U(0,1) \text{ 分布的随机数 } u_1,\cdots,u_n, \\ (2)\ \text{计算 } x_i = \begin{cases} 1, & u_i \leqslant p, \\ 0, & u_i > p, \end{cases} \quad i=1,\cdots,n, \\ (3)\ \text{计算 } x = \sum_{i=1}^{n} x_i. \end{cases}$$

该算法是十分简单的，但计算量大，需产生 n 个 $U(0,1)$ 分布的随机数并做 n 次比较.

(b)逆变换法. 应用逆变换法(6.21)可以产生 $b(n,p)$ 随机数，其中 $p_k = \binom{n}{k} p^k (1-p)^{n-k}$，$g_k = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$. 它只要产生一个 $U(0,1)$ 随机数，平均运行次数是 $1+np$.

(c)改进的逆变换法. 应用算法(6.23)也可以产生 $b(n,p)$ 随机数. 若取 m 为中位数，则可以算出它的平均运行次数为 $1.5+\sqrt{2/\pi}\sqrt{np(1-p)}$[4].

比较几种算法可见，(6.23)效果最好，它的平均运行次数与 $\sqrt{n}$ 成正比，而另两种算法的平均运行次数与 n 成正比. 当 n 很大时，为节约运行时间，也可采用正态近似来产生 $b(n,p)$ 随机数[4].

§ 6.1.6 随机向量的抽样方法

在用 Monte Carlo 等方法解应用问题时，随机向量的抽样也是经常用到的. 若随机向量各分量相互独立，则它等价于多个一元随机变量的抽样. 下面我们讨论非独立随机向量的抽样.

设 $X_1,\cdots,X_k$ 的概率密度函数为

$$p(x_1,\cdots,x_k)=p_1(x_1)p_2(x_2|x_1)\cdots p_k(x_k|x_1,\cdots,x_{k-1}) \tag{6.24}$$

其中 $p_1(x_1)$ 为 X_1 的边际分布密度函数，$p_i(x_i|x_1,\cdots,x_{i-1})$ 为给定 $X_1,\cdots,X_{i-1}$ 下 X_i 的条件分布密度函数.

定理 6.7 设 $U_1,\cdots,U_k$ 是独立同分布的 $U(0,1)$ 变量，$X_1,\cdots,X_k$ 是方程

$$\begin{cases}F(X_1)=U_1\\ F_i(X_i|X_1,\cdots,X_{i-1})=U_i, \quad i=2,\cdots,k\end{cases} \tag{6.25}$$

的解，其中 F_i 是对应于 p_i 的分布函数，则 $X_1,\cdots,X_k$ 的分布为(6.24).

证明：类似于定理 6.1 的证明，略.

定理 6.7 给出了随机向量的逆变换抽样方法：

$$\begin{cases}(1)\ 由\ U(0,1)\ 分布独立地抽取\ u_1,\cdots,u_k,\\ (2)\ 用方程(6.25)解\ x_1,\cdots,x_k.\end{cases} \tag{6.26}$$

由于联合密度函数的(6.24)表示不唯一(事实上有 $k!$ 种表示)，而不同的表示得到的方程(6.25)可能是不一样的，从而抽样的难易程度也不一样. 看一个例子.

例 6.8 设 X_1,X_2 的联合密度函数为

$$p(x_1,x_2)=\begin{cases}6x_1, & 如果\ x_1+x_2\leqslant 1,x_1\geqslant 0,x_2\geqslant 0\\ 0, & 其它\end{cases}$$

我们先将 $p(x_1,x_2)$ 写成 $p_1(x_1)p_2(x_2|x_1)$ 的形式. 此处 X_1 的边际分布密度函数为

$$p_1(x_1)=\int_0^{1-x_1}p(x_1,x_2)\mathrm{d}x_2=6x_1(1-x_1), \quad 0\leqslant x_1\leqslant 1$$

而给定 $X_1=x_1$ 下 X_2 的条件分布密度函数为

$$p_2(x_2|x_1)=\frac{p(x_1,x_2)}{p_1(x_1)}=\frac{1}{1-x_1}, \quad 0\leqslant x_2\leqslant 1-x_1$$

相应的边际分布函数和条件分布函数分别为

$$F_1(x_1) = \int_0^{x_1} p_1(t)\mathrm{d}t = 3x_1^2 - 2x_1^3, \quad 0 \leqslant x_1 \leqslant 1$$

$$F_2(x_2|x_1) = \int_0^{x_2} p_2(t|x_1)\mathrm{d}t = x_2(1-x_1)^{-1}, \quad 0 \leqslant x_2 \leqslant 1-x_1$$

方程(6.25)变为

$$\begin{cases} 3X_1^2 - 2X_1^3 = U_1 \\ X_2(1-X_1) = U_2 \end{cases} \tag{6.27}$$

下面我们再将 $p(x_1,x_2)$写成 $p_2(x_2)p_1(x_1|x_2)$的形式，其中

$$p_2(x_2) = \int_0^{1-x_2} p(x_1,x_2)\mathrm{d}x_1 = 3(1-x_2)^2, \quad 0 \leqslant x_2 \leqslant 1$$

$$p_1(x_1|x_2) = \frac{p(x_1,x_2)}{p_2(x_2)} = 2x_1(1-x_2)^{-2}, \quad 0 \leqslant x_1 \leqslant 1-x_2$$

相应的边际分布函数和条件分布函数分别为

$$F_2(x_2) = \int_0^{x_2} p_2(t)\mathrm{d}t = 1-(1-x_2)^2, \quad 0 \leqslant x_2 \leqslant 1$$

$$F_1(x_1|x_2) = \int_0^{x_1} p_1(t|x_2)\mathrm{d}t = x_1^2(1-x_2)^{-2}, \quad 0 \leqslant x_1 \leqslant 1-x_2$$

而方程(6.25)成为

$$\begin{cases} 1-(1-X_2)^2 = U_1 \\ X_1^2(1-X_2)^{-2} = U_2 \end{cases} \tag{6.28}$$

比较(6.27)与(6.28)，显然(6.28)要简单得多，它可以给出显式解：

$$X_2 = 1-(1-U_1)^{\frac{1}{3}}, \quad X_1 = (1-U_1)^{\frac{1}{3}}U_2^{\frac{1}{2}}$$

而解(6.27)则困难得多.

对服从特定的分布的随机向量有一些抽样方法，我们以例子的方式介绍多维正态变量的抽样.

例 6.9　设 $X=(X_1,\cdots,X_k)$服从 k 维正态分布 $N_k(\mu,\Sigma)$，其密度函数为

$$p(x) = (2\pi)^{-k/2}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)\right\}$$

此处 μ 为已知均值向量，$\boldsymbol{\Sigma}=(\sigma_{ij})_{k\times k}$ 为已知协差阵，$|\boldsymbol{\Sigma}|$ 是其行列式.

我们知道，因 $\boldsymbol{\Sigma}$ 是正定对称矩阵(此处不讨论奇异情形)，故存在下三角阵

$$\boldsymbol{C}=\begin{pmatrix} c_{11} & 0 & \cdots & 0 \\ c_{21} & c_{22} & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ c_{k1} & c_{k2} & \cdots & c_{kk} \end{pmatrix}$$

使得

$$\boldsymbol{\Sigma}=\boldsymbol{C}\boldsymbol{C}'$$

于是，若 $Z=(Z_1,\cdots,Z_k)\sim N_k(0,\boldsymbol{I}_k)$，则 $\boldsymbol{X}=\boldsymbol{C}Z+\mu\sim N_k(\mu,\boldsymbol{\Sigma})$.

Z 的抽样是直接的，关键是计算 $\boldsymbol{C}$，这可由迭代实现. 首先，由 $X_1=c_{11}Z_1+\mu_1$，有 $\sigma_{11}=\mathrm{Var}(X_1)=c_{11}^2$，从而 $c_{11}=\sqrt{\sigma_{11}}$. 接着考虑 X_2，因

$$X_2=c_{21}Z_1+c_{22}Z_2+\mu_2$$

于是，

$$\begin{aligned}\sigma_{22}&=\mathrm{Var}(X_2)=c_{21}^2+c_{22}^2\\ \sigma_{12}&=\mathrm{Cov}(X_1,X_2)\\ &=\mathrm{E}[c_{11}Z_1(c_{21}Z_1+c_{22}Z_2)]=c_{11}c_{21}\end{aligned}$$

从而给出

$$c_{21}=\frac{\sigma_{12}}{c_{11}}=\frac{\sigma_{12}}{\sqrt{\sigma_{11}}}$$

$$c_{22}=\left(\sigma_{22}-\frac{\sigma_{12}^2}{\sigma_{11}}\right)^{\frac{1}{2}}$$

如此进行，可得到一般的迭代公式：

$$c_{ij}=\frac{\sigma_{ij}-\sum_{l=1}^{j-1}c_{il}c_{jl}}{\left(\sigma_{jj}-\sum_{l=1}^{j-1}c_{jl}^2\right)^{\frac{1}{2}}},\qquad i=1,\cdots,k;\quad j=1,\cdots,i \tag{6.29}$$

至此，我们可以给出 k 维正态变量的抽样步骤：

$$\begin{cases}(1)\ \text{由}(6.29)\text{计算}\ c_{ij}, i=1,\cdots,k, j=1,\cdots,i,\\(2)\ \text{由}\ N(0,1)\ \text{分布独立抽取}\ k\ \text{个随机数}\ z_1,\cdots,z_k,\\(3)\ \text{计算}\ \boldsymbol{x}=\boldsymbol{C}\boldsymbol{z}+\boldsymbol{\mu},\text{其中},\boldsymbol{z}=(z_1,\cdots,z_k).\end{cases}$$

作为本节的结束,我们指出,在计算机上用数学方法产生$U(0,1)$随机数是一种得到广泛应用的方法.对此,由于$U(0,1)$分布的随机数的产生总是采用某个确定模型进行,从理论上讲总是会有周期现象出现的.而且,初值确定以后,所有随机数也都随之确定了,不满足真正随机数的要求.所以,通常把由数学方法产生的随机数称为伪随机数.然而,由于周期很长,在实际应用中几乎不可能出现.因而,人们还是把这种伪随机数当作真正的随机数处理.另外,我们也可对得到的随机数进行一系列检验以确定它们是否可以当作真正的随机数[1][2].

§6.2　随机模拟计算

掌握了各种分布变量的随机数的产生方法以后,我们就可以进行随机模拟计算了.本节首先介绍随机模拟计算的思想,然后主要就如何用模拟方法计算定积分进行讨论.

§6.2.1　统计模拟

随机模拟方法也称为 Monte Carlo 方法,其起源最早可以追溯到18世纪下半叶的 Buffon 试验.

例6.10　在1777年,法国学者 Buffon 提出用试验方法求圆周率π的值.其原理如下:假设平面上有无数条距离为1的等矩平行线,现向该平面随机地投掷一根长度为$l(l\leqslant 1)$的针,则我们可以计算该针与任一平行线相交的概率.此处随机投针可以这样理解:针的中心点与最近的平行线间的距离x均匀地分布在区间$[0,1/2]$上,针与平行线的夹角φ(不管相交与否)均匀地分布在区间$[0,\pi]$上(见图6.1).于是,针与线相交的充要条件是$\frac{x}{\sin\varphi}\leqslant\frac{l}{2}$,从而针线相交概率为

$$p \triangleq P\left(X\leqslant\frac{l}{2}\sin\varphi\right)=\int_0^{\pi}\int_0^{\frac{l}{2}\sin\varphi}\frac{2}{\pi}\mathrm{d}x\mathrm{d}\varphi=\frac{2l}{\pi}$$

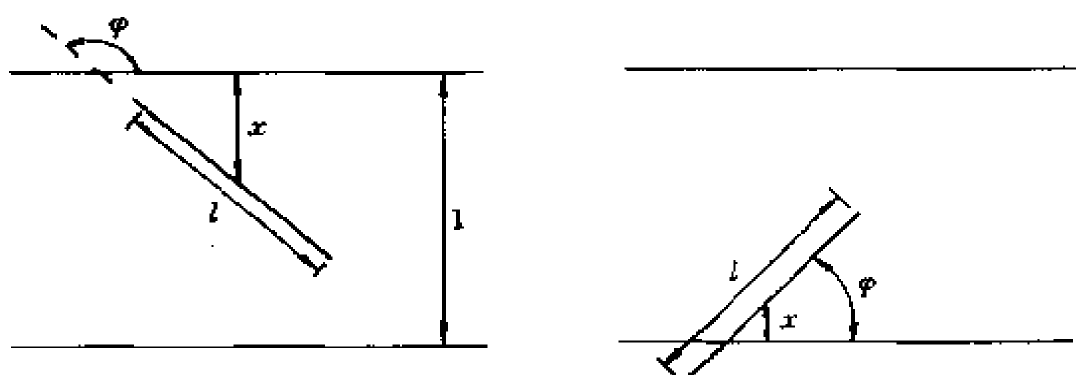


图 6.1 针与线的几种位置

根据以上公式，假如我们能作大量的投针试验并记录针与线相交的次数，则由大数定律可以估计出针线相交的概率 p，从而得到 π 的估计值. 历史上曾有几位学者做过这样的试验，其结果列于表 6.1.

表 6.1 Buffon 试验的结果

试验者	时间(年)	针长 l	投针次数	相交次数	π 估计值
Wolf	1850	0.80	5000	2532	3.159 56
Smith	1855	0.60	3204	1218	3.156 65
Fox	1884	0.75	1030	489	3.159 51
Lazzarini	1925	$0.8\dot{3}$	3408	1808	3.141 592 92

应该指出，上述试验的精度一般不会很高. 譬如，假设 $l=1$，则 $p=2/\pi$，由中心极限定理，若试验次数为 n，则 p 的估计 $\hat{p}$ 渐近服从 $N(p, p(1-p)/n)$，近似为 $N(0.6366, 0.2313/n)$. 因此，若要以 95% 的概率保证 $\hat{p}$ 精确到三位有效数字，即 $\hat{p}$ 与 p 的差距小于 0.001，则 n 必须满足 $n \geqslant 1.96^2 \times 0.2313/0.0001^2 \doteq 8.87 \times 10^5$. 表 6.1 中 Lazzarini 的试验次数不多而精度很高纯属巧合，因为它正好与祖冲之的密率355/113相等. 事实上，若该试验中的相交次数相差一次，都会对 π 的估计带来很大影响. 比如，若相交次数分别为 1807 和 1809，则 π 的估计分别为 3.143 331 和 3.139 856.

可以看出，上述试验是费时费力的，实施起来是困难的. 然而，随着计算机技术的快速发展，人们不需要具体实施这些试验，而只需要在计算机上进行模拟试验即可. 亦即是说，给定 l，我们可以在计算机上随机产生 x 和 φ，然后判断 $\frac{x}{\sin\varphi}\leqslant\frac{l}{2}$ 是否成立. 若成立，则针线相交，否则不交. 假如我们在计算机上独立地产生 10 000 对这样的 x 和 φ，并记录 $\frac{x}{\sin\varphi}\leqslant\frac{l}{2}$ 成立的次数，记为 n_0，则 π 的估计值可取为 $20\,000\times l/n_0$. 这就是随机模拟的计算结果.

由上述例子可以看出随机模拟计算的思路：

(1) 针对实际问题建立一个简单且便于实现的概率统计模型，使所求的解恰好是所建模型的概率分布或其某个数字特征，比如，是某个事件的概率，或者是该模型的期望值；

(2) 对模型中的随机变量建立抽样方法，在计算机上进行模拟试验，抽取足够的随机数，并对有关的事件进行统计；

(3) 对模拟试验结果加以分析，给出所求解的估计及其精度(方差)的估计；

(4) 必要时，还应改进模型以降低估计方差和减少试验费用，提高模拟计算的效率.

在上述模拟试验和分析中，涉及到各种抽样方法和分析方法，因此，这里面必然也有结果好坏的差别. 本节后面将对此加以讨论，介绍一些改进结果精度的方法.

Monte Carlo 方法近年来应用十分广泛，我们在后面几节介绍的一些动态 Monte Carlo 方法更是近年来发展最快的统计分支之一. 本节先介绍定积分的 Monte Carlo 计算.

考虑一个简单的定积分

$$\theta = \int_a^b f(x)\mathrm{d}x \tag{6.30}$$

不少统计问题，如计算概率、各阶矩等，最后都归结为定积分的近似计算问题，这里我们首先介绍两种求 θ 的简单的 Monte Carlo 方法，然后给出一些较复杂而更有效的 Monte Carlo 方法.

§6.2.2 随机投点法

考虑(6.30)的积分. 简单起见，设 a,b 有限，$0\leqslant f(x)\leqslant M$，令 $\Omega=\{(x,y): a\leqslant x\leqslant b,\ 0\leqslant y\leqslant M\}$，并设 (X,Y) 是在 Ω 上均匀分布的二维随机变量，其联合密度函数为 $\frac{1}{M(b-a)}I_{(a\leqslant x\leqslant b,0\leqslant y\leqslant M)}$. 则易见 $\theta=\int_a^b f(x)\mathrm{d}x$ 是 Ω 中曲线 $y=f(x)$ 下方的面积(见图 6.2). 假设我们向 Ω

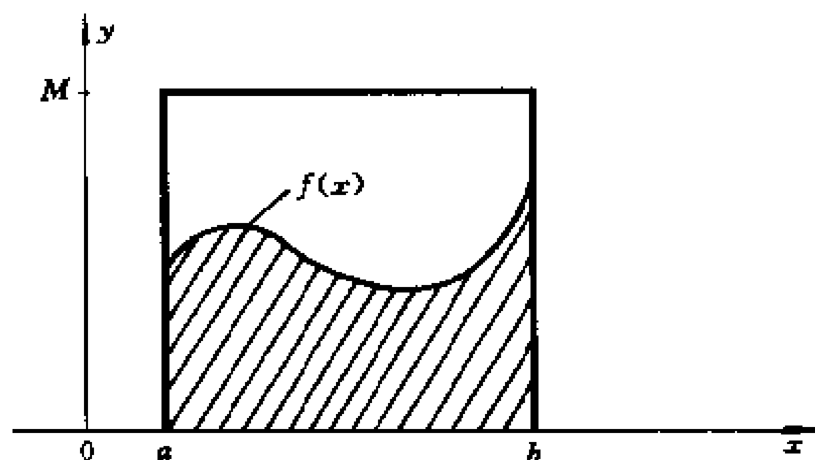


图 6.2

中进行随机投点，若点落在 $y=f(x)$ 下方(即 $y<f(x)$)称为中的，否则称为不中，则点中的的概率为 $p=\theta/M(b-a)$. 若我们进行了 n 次投点，其中 n_0 次中的，则可以得到 θ 的一个估计

$$\hat{\theta}_1=M(b-a)\frac{n_0}{n} \tag{6.31}$$

上述思想是容易实施的. 我们只要独立地产生 n 对 (u_i,v_i)，其中 u_i,v_i 是来自 $U(0,1)$ 的独立随机数，令 $x_i=a+u_i(b-a)$，$y_i=Mv_i$，然后看是否有 $f(x_i)\geqslant y_i$. 若 $f(x_i)\geqslant y_i$，则中的，如此可计算出中的次数 n_0. 具体步骤为

$$\begin{cases}(1)\ 独立地产生\ 2n\ 个\ U(0,1)\ 随机数, u_i,v_i,i=1,\cdots,n,\\(2)\ 计算\ x_i=a+u_i(b-a),y_i=Mv_i,和\ f(x_i),\\(3)\ 统计\ f(x_i)\geqslant y_i\ 的个数\ n_0,\\(4)\ 用(6.31)\ 估计\ \theta.\end{cases} \tag{6.32}$$

随机投点法的想法简单明了，而且，因每次投点结果服从二点分布，故 $n_0 \sim b(n, p)$，其中 $p = \theta / M(b-a)$. 不难看出，$\hat{\theta}_1$ 是 θ 的无偏估计，且其方差为

$$\mathrm{Var}(\hat{\theta}_1) = \frac{M^2(b-a)^2}{n^2}\mathrm{Var}(n_0)$$

$$= \frac{\theta}{n}[M(b-a) - \theta] \tag{6.33}$$

由(6.33)，若用估计的标准差来衡量其精度，则估计 $\hat{\theta}_1$ 的精度的阶为 $n^{-1/2}$.

§6.2.3　样本平均值法

对积分 $\theta = \int_a^b f(x)\mathrm{d}x$，设 $g(x)$ 是 (a, b) 上的一个密度函数，改写

$$\theta = \int_a^b \frac{f(x)}{g(x)} g(x)\mathrm{d}x = \mathrm{E}\left[\frac{f(X)}{g(X)}\right]$$

可见，任一积分均可以表示为某个随机变量(函数)的期望. 由矩法，若有 n 个来自 $g(x)$ 的观测值，则可给出 θ 的一个矩估计，这便是样本平均值法的基本原理.

最简单地，若 a, b 有限，可取 $g(x) = 1/(b-a)$. 设 $x_1, \cdots, x_n$ 是来自 $U(a, b)$ 的随机数，则 θ 的一个估计为

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}\frac{f(x_i)}{g(x_i)} = \frac{b-a}{n}\sum_{i=1}^{n} f(x_i) \tag{6.34}$$

其具体计算步骤为

$$\begin{cases} (1)\ \text{独立地产生}\ n\ \text{个}\ U(0,1)\ \text{随机数}\ u_1, \cdots, u_n, \\ (2)\ \text{计算}\ x_i = a + (b-a)u_i\ \text{和}\ f(x_i), i = 1, \cdots, n, \\ (3)\ \text{用(6.34)估计}\ \theta. \end{cases} \tag{6.35}$$

显然，$\hat{\theta}_2$ 是 θ 的无偏估计. 我们也可计算其方差.

$$\mathrm{Var}(\hat{\theta}_2) = \mathrm{Var}\left[\frac{1}{n}(b-a)\sum_{i=1}^{n} f(X_i)\right]$$

$$= \frac{1}{n}\left[(b-a)^2\int_a^b (f(x))^2 \frac{1}{b-a}\mathrm{d}x - \theta^2\right] \tag{6.36}$$

$$= \frac{1}{n}\left[(b-a)\int_a^b f^2(x)\mathrm{d}x - \theta^2\right]$$

在假设 $0 \leqslant f(x) \leqslant M$ 下，可以证明，对相同的 n，$\mathrm{Var}(\hat{\theta}_2) \leqslant \mathrm{Var}(\hat{\theta}_1)$. 事实上

$$\begin{aligned}\mathrm{Var}(\hat{\theta}_1) - \mathrm{Var}(\hat{\theta}_2) &= \frac{1}{n}\theta[M(b-a)-\theta] - \\ &\quad \frac{1}{n}\left[(b-a)\int_a^b f^2(x)\mathrm{d}x - \theta^2\right] \\ &= \frac{M(b-a)}{n}\left[\theta - \int_a^b \frac{f^2(x)}{M}\mathrm{d}x\right] \\ &\geqslant \frac{M(b-a)}{n}\left[\theta - \int_a^b f(x)\mathrm{d}x\right] = 0\end{aligned}$$

而且，只要 $\{x: f(x) < M\}$ 不是零测集，就有 $\mathrm{Var}(\hat{\theta}_2) < \mathrm{Var}(\hat{\theta}_1)$. 由此，样本平均值法比随机投点法要有效.

从上述介绍可以看出，样本平均值法并不要求一定有 $f(x) \leqslant M$，它可推广至一般情形.

§6.2.4 重要抽样方法(Importance Sample)

我们已经看到，随机投点法和样本平均值法都可以估计(近似计算)θ，但二者的效(方差)不同. 自然，人们会进一步考虑这样一个问题：是不是有更好的(Monte Carlo)估计方法，它的方差比 $\hat{\theta}_2$ 的方差还要小. 这是 Monte Carlo 方法中一类重要的研究课题，即考虑一些降低估计方差的技术. 本小节开始，我们介绍几种这样的技术. 首先介绍一种最常用的降低方差的方法：重要抽样法.

自样本平均值法知，对任一密度函数 $g(x)$，$\theta = \int_a^b \frac{f(x)}{g(x)} g(x)\mathrm{d}x = \mathrm{E}\left[\frac{f(X)}{g(X)}\right]$. 因此，若有来自 $g(x)$ 的样本(随机数)$x_1, \cdots, x_n$，则 θ 的估计可取为

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(x_i)}{g(x_i)} \tag{6.37}$$

$\hat{\theta}$ 显然是无偏的，其方差则与 $g(\cdot)$ 有关. 问题变为，如何选择 $g(\cdot)$ 使

得 $\hat{\theta}$ 的方差小.

从理论上看,因 $\mathrm{Var}(\hat{\theta})=\frac{1}{n}\left[\mathrm{E}\left(\frac{f(X)}{g(X)}\right)^2-\theta^2\right]$,若 $f(x)\geqslant 0$,取 $g(x)=f(x)/\theta$,则有 $\mathrm{Var}(\hat{\theta})=0$. 因为 θ 未知,这个结果是无用的. 然而,虽然我们不能取 $g(x)=f(x)/\theta$,但它也给我们一个启示,亦即,$g(x)$与 $f(x)$形状接近应能降低估计的方差. 这便是重要抽样法的基本思想.

直观地看,(6.34)式给出的 $\hat{\theta}_2$ 是采用均匀抽样,诸 x_i 是均匀分布的随机数,各 x_i 对 $\hat{\theta}_2$ 的贡献是不同的,$f(x_i)$大则贡献大,但在抽样时这种差别未能体现出来,因此效不会很高,要达到同样的精度需要的抽样次数多. 而重要抽样法则希望贡献率大的随机数出现的概率大,贡献小的随机数出现概率小,从而提高抽样的效.

例 6.11　考虑一个简单的积分 $\theta=\int_0^1 \mathrm{e}^x\mathrm{d}x$. 此处 θ 的精确值($\mathrm{e}-1$)是可求的. 为说明重要抽样法,我们采用 Monte Carlo 方法估计 θ.

首先考虑样本平均值法,即,产生 n 个 $U(0,1)$随机数 $x_1,\cdots,x_n$,则 $\hat{\theta}=\frac{1}{n}\sum_{i=1}^n \mathrm{e}^{x_i}$,且 $\mathrm{E}(\hat{\theta})=\theta$,

$$
\begin{aligned}
\mathrm{Var}(\hat{\theta}) &= \frac{1}{n}\left[\int_0^1 \mathrm{e}^{2x}\mathrm{d}x-(\mathrm{e}-1)^2\right] \\
&= \frac{1}{n}\left[\frac{1}{2}(\mathrm{e}^2-1)-(\mathrm{e}-1)^2\right]=\frac{0.242}{n}
\end{aligned}
$$

由重要抽样法的思想,我们要选一个与 e^x 相似的密度函数. 由 Taylor 展开式,$\mathrm{e}^x=1+x+\frac{x^2}{2}+\cdots$. 利用线性近似,取 $g(x)=\frac{2}{3}(1+x)$,则 $g(x)$是(0,1)上的一个密度函数. 设 $x_1,\cdots,x_n$ 是来自 $g(x)$的随机数,则 θ 的估计可取为

$$
\tilde{\theta}=\frac{1}{n}\sum_{i=1}^n\frac{f(x_i)}{g(x_i)}=\frac{3}{2n}\sum_{i=1}^n\frac{\mathrm{e}^{x_i}}{1+x_i}
$$

$\tilde{\theta}$ 也是 θ 的无偏估计,且

$$
\mathrm{Var}(\tilde{\theta})=\frac{1}{n}\left[\int_0^1\frac{f^2(x)}{g(x)}\mathrm{d}x-(\mathrm{e}-1)^2\right]
$$

$$= \frac{1}{n}\left[\int_0^1 \frac{3}{2}\,\frac{e^{2x}}{1+x}\mathrm{d}x - (e-1)^2\right]$$

$$= \frac{0.0269}{n}（数值计算）$$

由此，$\tilde{\theta}$ 明显优于 $\hat{\theta}$.

§ 6.2.5　分层抽样方法

另一种利用贡献率大小来降低估计方差的方法是分层抽样法. 它首先把样本空间 D 分成一些小区间 $D_1,\cdots,D_m$，且诸 D_i 不交，$\bigcup D_i = D$，然后在各小区间内的抽样数由其贡献大小决定，亦即，定义 $p_i = \int_{D_i} f(x)\mathrm{d}x$，则，$D_i$ 内的抽样数应与 p_i 成正比. 如此，对 θ 贡献大的 D_i 抽样多，可提高抽样效率.

考虑积分 $\theta = \int_0^1 f(x)\mathrm{d}x$，将[0,1]分成 m 个小区间，各区间端点记为 a_i，$0=a_0<a_1<\cdots<a_m=1$，则

$$\theta = \int_0^1 f(x)\mathrm{d}x = \sum_{i=1}^{m}\int_{a_{i-1}}^{a_i} f(x)\mathrm{d}x \triangleq \sum_{i=1}^{m} I_i$$

记 $l_i = a_i - a_{i-1}$ 为第 i 个小区间的长度，$i=1,\cdots,m$，在每个小区间上的积分值 I_i 可用样本平均值法估计出来，然后将之相加即可给出 θ 的一个估计. 具体步骤为：

$$\begin{cases}
(1)\ 产生\ U(0,1)\ 随机数\{u_{ij}, j=1,\cdots,n_i, i=1,\cdots,m\}, \\
(2)\ 计算\ x_{ij} = a_{i-1} + l_i u_{ij}, j=1,\cdots,n_i, i=1,\cdots,m, \\
(3)\ 计算\ \hat{I}_i = \dfrac{l_i}{n_i}\sum_{j=1}^{n_i} f(x_{ij}).
\end{cases}$$

于是可得到 θ 的估计为

$$\hat{\theta}_3 = \sum_{i=1}^{m}\hat{I}_i \tag{6.38}$$

由样本平均值法，$\mathrm{E}\hat{I}_i = I_i$，因而 $\hat{\theta}_3$ 是 θ 的无偏估计，其方差为

$$\mathrm{Var}(\hat{\theta}_3) = \mathrm{Var}\left\{\sum_{i=1}^{m}\frac{l_i}{n_i}\sum_{j=1}^{n_i} f(X_{ij})\right\}$$

$$= \sum_{i=1}^{m} \frac{l_i^2}{n_i} \sigma_i^2 \tag{6.39}$$

其中

$$\sigma_i^2 = \int_{a_{i-1}}^{a_i} \frac{f^2(x)}{l_i} \mathrm{d}x - \left(\frac{I_i}{l_i}\right)^2 \tag{6.40}$$

例 6.12　再看例 6.11,我们先将抽样区间[0,1]划分为两个小区间:[0,0.5)和[0.5,1],则

$$I_1 = \int_0^{0.5} \mathrm{e}^x \mathrm{d}x = \sqrt{\mathrm{e}} - 1$$

$$I_2 = \int_{0.5}^{1} \mathrm{e}^x \mathrm{d}x = \mathrm{e} - \sqrt{\mathrm{e}}$$

$$\sigma_1^2 = \int_0^{0.5} 2\mathrm{e}^{2x} \mathrm{d}x - 4(\sqrt{\mathrm{e}} - 1)^2$$

$$= (\mathrm{e} - 1) - 4(\sqrt{\mathrm{e}} - 1)^2 = 0.034\,92$$

$$\sigma_2^2 = \int_{0.5}^{1} 2\mathrm{e}^{2x} \mathrm{d}x - 4(\mathrm{e} - \sqrt{\mathrm{e}})^2$$

$$= (\mathrm{e}^2 - \mathrm{e}) - 4(\mathrm{e} - \sqrt{\mathrm{e}})^2 = 0.094\,93$$

设一共抽 n 个随机数,其中在[0,0.5)上抽 n_1 个,则使用分层抽样求得的 $\hat{\theta}_3$ 的方差为

$$\mathrm{Var}(\hat{\theta}_3) = \frac{0.5^2}{n_1}\sigma_1^2 + \frac{0.5^2}{n - n_1}\sigma_2^2$$

对 n_1 求导,易知在 n 固定下,当

$$\frac{n_1}{n} = \frac{\sigma_1}{\sigma_1 + \sigma_2} = \frac{\sqrt{0.034\,92}}{\sqrt{0.034\,92} + \sqrt{0.094\,93}}$$

$$= \frac{0.186\,87}{0.186\,87 + 0.308\,11} = 0.377\,53$$

时,$\hat{\theta}_3$ 的方差最小,为

$$\mathrm{Var}(\hat{\theta}_3) = \frac{0.5^2}{n}\left[\frac{\sigma_1^2}{n_1/n} + \frac{\sigma_2^2}{1 - n_1/n}\right] = \frac{0.061\,25}{n}$$

与例 6.11 的结果比较,有 $\frac{0.0269}{n} < \frac{0.061\,25}{n} < \frac{0.242}{n}$.

如果我们将区间进一步分细,比如将[0,1]10 等分,我们也可计算

诸 σ_i^2，并确定最优的抽样次数分配 $n_i/n=\sigma_i\Big/\sum_{j=1}^{m}\sigma_j$结果见表 6.2，从而给出分层抽样法估计的方差为0.002 46/n，它远远小于 $\hat{\theta}$ 的方差 0.0269/n.

表 6.2 10 等分[0,1]下诸 σ_i^2 和 n_i

i	1	2	3	4	5
σ_i^2	0.0009	0.0011	0.0014	0.0017	0.0021
σ_i	0.0304	0.0336	0.0371	0.0410	0.0453
n_i/n	0.0612	0.0676	0.0748	0.0826	0.0913
i	6	7	8	9	10
σ_i^2	0.0025	0.0031	0.0037	0.0046	0.0056
σ_i	0.0501	0.0553	0.0611	0.0676	0.0747
n_i/n	0.1009	0.1115	0.1233	0.1362	0.1505

分层抽样法可以降低估计的方差，而且，如果我们将抽样区间分得足够小，并较好地分配抽样次数，则可使估计方差大大降低. 我们已指出，分层抽样法的估计方差为(6.39)式，若诸 σ_i^2 和 l_i 已知，则不难证明，在 n 固定下，当

$$n_i = nl_i\sigma_i/\Big(\sum_{i=1}^{m}l_i\sigma_i\Big) \tag{6.41}$$

时，估计的方差达最小，其最小值为$\frac{1}{n}(\sum_{i=1}^{m}l_i\sigma_i)^2$.

分层抽样法有两个主要问题，其一是怎样划分抽样区间，简单而常用的方法是将抽样区间等分；另一个问题是，在抽样区间划分好以后，如何确定抽样次数的分配. (6.41)给出了一个最优分配，然而在实际中，由于 σ_i^2 总是未知的，因而(6.41)无法应用. 即使如此，分层抽样法还是有其作用的. 我们指出，即使取简单的分配：$n_i=nl_i\Big/\sum l_i=nl_i/(b-a)$，也有 $\mathrm{Var}(\hat{\theta}_3)\leqslant\mathrm{Var}(\hat{\theta}_2)$.

事实上，我们以 $n_i=nl_i/(b-a)$代入(6.39)，有

$$\mathrm{Var}(\hat{\theta}_3) = \frac{b-a}{n}\sum_{i=1}^{m} l_i \sigma_i^2$$

由 Cauchy-Schwarz 不等式，有

$$\theta^2 = \left(\sum_{i=1}^{m} I_i\right)^2 = \left[\sum_{i=1}^{m} \frac{I_i}{\sqrt{l_i}}\sqrt{l_i}\right]^2$$

$$\leqslant \sum_{i=1}^{m} \frac{I_i^2}{l_i}\sum_{i=1}^{m} l_i = (b-a)\sum_{i=1}^{m} \frac{I_i^2}{l_i}$$

据此，在(6.40)式中两端各乘以 l_i 并相加，即有

$$\sum_{i=1}^{m} l_i\sigma_i^2 = \int_a^b f^2(x)\mathrm{d}x - \sum_{i=1}^{m} \frac{I_i^2}{l_i}$$

$$\leqslant \int_a^b f^2(x)\mathrm{d}x - \theta^2/(b-a)$$

这就给出了 $\mathrm{Var}(\hat{\theta}_3) \leqslant \mathrm{Var}(\hat{\theta}_2)$.

§6.2.6　关联抽样方法

为引出关联抽样法，我们考虑下面的积分之差

$$\theta = \int f_1(x)\mathrm{d}x - \int f_2(x)\mathrm{d}x \triangleq I_1 - I_2$$

假设我们分别用 $\hat{I}_1$ 和 $\hat{I}_2$ 估计 I_1 和 I_2，并用 $\hat{\theta}=\hat{I}_1-\hat{I}_2$ 估计 θ，则 $\hat{\theta}$ 的方差为

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Var}(\hat{I}_1) + \mathrm{Var}(\hat{I}_2) - 2\mathrm{Cov}(\hat{I}_1, \hat{I}_2)$$

显然，在 $\mathrm{Var}(\hat{I}_1)$ 和 $\mathrm{Var}(\hat{I}_2)$ 确定后，$\hat{I}_1$ 与 $\hat{I}_2$ 的正相关度越高，则 $\hat{\theta}$ 的方差越小. 这便是关联抽样法的基本出发点.

考虑用重要抽样法估计 I_1 和 I_2，即改写 θ 为

$$\theta = \int h_1(x)g_1(x)\mathrm{d}x - \int h_2(x)g_2(x)\mathrm{d}x$$

其中 $g_1(x), g_2(x)$ 是两个密度函数，$h_i(x)=f_i(x)/g_i(x)$，$i=1,2$. 首先，我们由 $g_1(x), g_2(x)$ 各产生 n 个随机数 $X_1,\cdots,X_n$ 和 $Y_1,\cdots,Y_n$，然后计算

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{m}(h_1(X_i) - h_2(Y_i))$$

如果我们设计 X_i 与 Y_i 有较高的正相关，比如，采用逆变换方法由同一个 $U(0,1)$ 分布的随机数产生 X_i 与 Y_i，则可望降低估计的方差.

用类似的想法我们可以改进 $\theta=\int f(x)\mathrm{d}x$ 的估计. 首先讨论 $\theta=\int_0^1 f(x)\mathrm{d}x$ 的情况，因

$$\int_0^1 f(x)\mathrm{d}x=\int_0^1 f(1-x)\mathrm{d}x$$

故

$$\theta=\frac{1}{2}\int_0^1[f(x)+f(1-x)]\mathrm{d}x\triangleq I_1+I_2$$

于是，我们可先产生 n 个 $U(0,1)$ 分布的随机数 $u_1,\cdots,u_n$，然后计算

$$\hat{\theta}_4=\frac{1}{2n}\sum_{i=1}^{n}[f(u_i)+f(1-u_i)] \tag{6.42}$$

我们可将 $\hat{\theta}_4$ 与(6.34)式的 $\hat{\theta}_2$ 进行比较. 此处 $\hat{\theta}_4$ 显然也是无偏的，且有下面的定理。

定理 6.8 在上述假定下，若 $f(x)$ 是单调函数且具有一阶连续导数，则

$$\mathrm{Var}(\hat{\theta}_4)\leqslant\frac{1}{2}\mathrm{Var}(\hat{\theta}_2)$$

证明：不失一般性，设 $n=1$，则

$$\begin{aligned}\mathrm{Var}(\hat{\theta}_4)&=\frac{1}{4}\left[\int_0^1(f(x)+f(1-x))^2\mathrm{d}x-(2\theta)^2\right]\\&=\frac{1}{2}\int_0^1 f^2(x)\mathrm{d}x+\frac{1}{2}\int_0^1 f(x)f(1-x)\mathrm{d}x-\theta^2\end{aligned}$$

又由(6.36)式($a=0,b=1,n=1$)，$\mathrm{Var}(\hat{\theta}_2)=\int_0^1 f^2(x)\mathrm{d}x-\theta^2$，于是

$$2\mathrm{Var}(\hat{\theta}_4)-\mathrm{Var}(\hat{\theta}_2)=\int_0^1 f(x)f(1-x)\mathrm{d}x-\theta^2 \tag{6.43}$$

由定理条件，$f(x)$ 为单调函数，不妨设为单调增函数，于是 $f(1)\geqslant\theta\geqslant f(0)$. 令

$$\varphi(t)=\int_0^t f(1-x)\mathrm{d}x-t\theta$$

则 $\varphi(0)=\varphi(1)=0$，且 $\varphi'(t)=f(1-t)-\theta$ 是单调减函数. 而 $\varphi'(0)=f(1)-\theta\geqslant 0$，$\varphi'(1)=f(0)-\theta\leqslant 0$，从而有 $\varphi(x)\geqslant 0$，$\forall\ x\in[0,1]$. 由此，

$$\int_0^1\varphi(x)f'(x)\mathrm{d}x\geqslant 0$$

将上式分部积分，给出

$$\int_0^1\varphi'(x)f(x)\mathrm{d}x\leqslant 0$$

亦即

$$\int_0^1 f(1-x)f(x)\mathrm{d}x-\theta\int_0^1 f(x)\mathrm{d}x\leqslant 0$$

将之代入(6.43)式，定理得证.

更一般地，考虑 $\theta=\int_a^b h(x)g(x)\mathrm{d}x$，其中 $g(x)$ 是一密度函数，$h(x)$ 相当于前述 $f(x)/g(x)$. 由与导出(6.42)式类似的想法，我们可给出

$$\hat{\theta}_4=\frac{1}{2n}\sum_{i=1}^{n}[h(x_i)+h(x'_i)]$$

其中 $x_i=G^{-1}(u_i)$，$x'_i=G^{-1}(1-u_i)$，$u_1,\cdots,u_n$ 为来自 $U(0,1)$ 的随机数，G 为 g 对应的分布函数. 由构造，(X_i,X'_i) 是负相关的. 若 $h(x)$ 单调，则 $\hat{\theta}_4$ 的方差比样本平均值法得到的估计的方差小.

我们上面的讨论均在假定 a,b 有限下进行，读者不难发现，当 a,b 为无穷时，上述方法照样可行. 以样本平均值法为例. 设 $\theta=\int_0^{\infty}f(x)\mathrm{d}x$，取 $g(x)=\mathrm{e}^{-x}$(标准指数分布)，令 $x_1,\cdots,x_n$ 是来自 $g(x)$ 的随机数，则 θ 的一个估计是

$$\hat{\theta}=\frac{1}{n}\sum_{i=1}^{n}f(x_i)\mathrm{e}^{x_i}$$

§ 6.3 EM 算法及其推广

在统计领域里，主要有两大类计算问题，一类是极大似然估计的计算，另一类是 Bayes 计算. 从计算方法上讲，这二者是可以合并讨论的，

因极大似然估计的计算类似于 Bayes 的后验众数的计算，因此，我们后面就从 Bayes 计算角度介绍统计计算方法.

Bayes 计算方法已有很多，大体上可分为两大类. 一类是直接应用于后验分布以得到后验均值或后验众数的估计，以及这种估计的渐近方差或其近似. 譬如上一节我们介绍的定积分的统计模拟计算，它只能应用于比较简单的后验分布. 另一类算法可以总称为数据添加算法，这是近年发展很快且应用很广的一种算法，它不是直接对复杂的后验分布进行极大化或进行模拟，而是在观测数据的基础上加上一些"潜在数据"，从而简化计算并完成一系列简单的极大化或模拟，该"潜在数据"可以是"缺损数据(Missing Data)"或未知参数. 其原理可表述如下：设我们能观测到的数据是 Y,θ 关于 Y 的后验分布 $p(\theta|Y)$ 很复杂，难以直接进行各种统计计算，假如我们能假定一些没有能观测到的潜在数据 Z 为已知(譬如，Y 为某变量的截尾观测值，而 Z 为该变量的真值)，则可能得到一个关于 θ 的简单的添加后验分布 $p(\theta|Y,Z)$，利用 $p(\theta|Y,Z)$ 的简单性我们可进行各种统计计算. 如极大化，抽样等，然后回过头来，我们又可以对 Z 的假定作检查和改进，如此进行，我们就将一个复杂的极大化或抽样问题转变为一系列简单的极大化或抽样. 当然，这里面有一个收敛性的问题，涉及到方法的选取. 我们下面介绍两种常用的数据添加算法，它们是 EM 算法及其各种推广和 Markov Chain Monte Carlo 方法. 本节先介绍 EM 算法.

§6.3.1 EM 算法

先考虑一个简单情形. 设某元件的失效时间 Y 关于某个变量 x 有直线回归关系，假设在一次试验中得到一批观测数据，见图 6.3，叉"×"表示该种元件的失效时间对应的值，小圆圈"∘"对应元件的(右)截尾时间(比实际失效时间要小). 如果直线斜率和截距的估计值已知，则我们可以在真实数据不小于截尾数据的前提下，将各个被截尾的失效时间估计出来(譬如，若 $\mathrm{E}(Y|x)>Z$，则用 $\mathrm{E}(Y|x)$ 作为真实数据，否则取 Z 作为真实数据)，从而得到所谓的"完全数据"，由此"完全数据"，我们可以对直线斜率和截距进行估计(如极大似然估计)，估计出

新的斜率和截距后，再重新估计各个被截尾的失效时间，得到新的完全数据，如此重复，我们便将一个复杂的估计问题替换成一系列简单的估计问题，关于它的收敛性，以及这样得到的斜率和截距的估计是否具有一些好的性质，我们将在后面讨论．将之一般化，我们可给出 EM 算法．

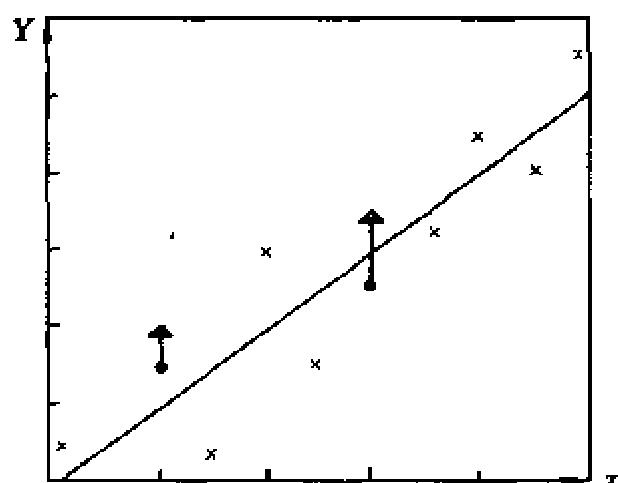


图 6.3　试验数据

EM 算法是一种迭代方法，最初由 Dempster 等提出[6]，并主要用来求后验分布的众数(即极大似然估计)，它的每一次迭代由两步组成：E 步(求期望)和 M 步(极大化)．一般地，以 $p(\theta|Y)$表示 θ 的基于观测数据的后验分布密度函数，称为观测后验分布，$p(\theta|Y,Z)$表示添加数据 Z 后得到的关于 θ 的后验分布密度函数，称为添加后验分布，$p(Z|\theta,Y)$表示在给定 θ 和观测数据 Y 下潜在数据 Z 的条件分布密度函数．我们的目的是计算观测后验分布 $p(\theta|Y)$的众数，于是，EM 算法如下进行．记 $\theta^{(i)}$ 为第 $i+1$ 次迭代开始时后验众数的估计值，则第 $i+1$ 次迭代的两步为

E 步：将 $p(\theta|Y,Z)$或 $\log p(\theta|Y,Z)$关于 Z 的条件分布求期望，从而把 Z 积掉，即

$$\begin{aligned} Q(\theta|\theta^{(i)},Y) &\triangleq E_Z[\log p(\theta|Y,Z)|\theta^{(i)},Y] \\ &= \int \log[p(\theta|Y,Z)]p(Z|\theta^{(i)},Y)\mathrm{d}Z \qquad (6.44) \end{aligned}$$

M 步：将 $Q(\theta|\theta^{(i)},Y)$ 极大化，即找一个点 $\theta^{(i+1)}$，使

$$Q(\theta^{(i+1)}|\theta^{(i)},Y)=\max_{\theta}Q(\theta|\theta^{(i)},Y) \tag{6.45}$$

如此形成了一次迭代 $\theta^{(i)}\to\theta^{(i+1)}$. 将上述 E 步和 M 步进行迭代直至 $\|\theta^{(i+1)}-\theta^{(i)}\|$ 或 $\|Q(\theta^{(i+1)}|\theta^{(i)},Y)-Q(\theta^{(i)}|\theta^{(i)},Y)\|$ 充分小时停止.

下面我们看一个例子.

例 6.13　假设一次试验可能有四个结果，其发生的概率分别为 $\frac{1}{2}+\frac{\theta}{4},\frac{1}{4}(1-\theta),\frac{1}{4}(1-\theta),\frac{\theta}{4}$，其中 $\theta\in(0,1)$，现进行了 197 次试验，四种结果的发生次数分别为 125，18，20，34. 此处观测数据为

$$Y=(y_1,y_2,y_3,y_4)=(125,18,20,34)$$

取 θ 的先验分布 $\pi(\theta)$ 为平坦分布（此处即 $(0,1)$ 上均匀分布），则 θ 的观测后验分布为

$$\begin{aligned}p(\theta|Y)&\propto\pi(\theta)p(Y|\theta)\\&=\left(\frac{1}{2}+\frac{\theta}{4}\right)^{y_1}\left[\frac{1}{4}(1-\theta)\right]^{y_2}\left[\frac{1}{4}(1-\theta)\right]^{y_3}\left(\frac{\theta}{4}\right)^{y_4}\\&\propto(2+\theta)^{y_1}(1-\theta)^{y_2+y_3}\theta^{y_4}\end{aligned} \tag{6.46}$$

现假设第一种结果可以分解成两部分，其发生概率分别为 1/2 和 $\theta/4$，令 Z 和 y_1-Z 分别表示试验中结果落入这两部分的次数（Z 是不能观测的潜在数据），则 θ 的添加后验分布为

$$\begin{aligned}p(\theta|Y,Z)&\propto\pi(\theta)p(Y,Z|\theta)\\&=\left(\frac{1}{2}\right)^{z}\left(\frac{\theta}{4}\right)^{y_1-z}\left[\frac{1}{4}(1-\theta)\right]^{y_2}\left[\frac{1}{4}(1-\theta)\right]^{y_3}\left(\frac{\theta}{4}\right)^{y_4}\\&\propto\theta^{y_1-z+y_4}(1-\theta)^{y_2+y_3}\end{aligned} \tag{6.47}$$

用(6.46)求 θ 的后验众数比较麻烦，而用(6.47)式求后验众数则十分简单，因此，我们可用 EM 算法求(6.46)式的后验众数.

我们在(6.46)和(6.47)中均采用了符号 $\propto$，它表示该符号两端可能存在一个与 θ 无关的比例因子. 这个比例因子在 EM 算法中是不起作用的，因为它在极大化(6.45)时可约去. 后面我们在许多地方均要采用符号 $\propto$，它使许多计算简化.

在第 $i+1$ 次迭代中，假设有估计值 $\theta^{(i)}$，则可通过 E 步和 M 步得

到 θ 的一个新的估计. 在 E 步中,由(6.44)有

$$Q(\theta|\theta^{(i)},Y)=\mathrm{E}^{Z}[(y_1-Z+y_4)\log\theta+(y_2+y_3)\log(1-\theta)|\theta^{(i)},Y]$$

$$=[y_1-\mathrm{E}^{Z}(Z|\theta^{(i)},Y)+y_4]\log\theta+(y_2+y_3)\log(1-\theta)$$

因在 $\theta^{(i)}$ 和 Y 给定下,$Z\sim b(y_1,2/(\theta^{(i)}+2))$,故

$$\mathrm{E}^{Z}(Z|\theta^{(i)},Y)=2y_1/(\theta^{(i)}+2)$$

在 M 步中,我们将 $Q(\theta|\theta^{(i)},Y)$ 对 θ 求导并令其为 0,有

$$\theta^{(i+1)}=\frac{y_1+y_4-\mathrm{E}^{Z}(Z|\theta^{(i)},Y)}{y_1+y_2+y_3+y_4-\mathrm{E}^{Z}(Z|\theta^{(i)},Y)}$$

$$=\frac{\theta^{(i)}y_1+(\theta^{(i)}+2)y_4}{\theta^{(i)}y_1+(\theta^{(i)}+2)(y_2+y_3+y_4)}$$

$$=\frac{159\theta^{(i)}+68}{197\theta^{(i)}+144} \tag{6.48}$$

(6.48)式给出了由 EM 算法得到的迭代公式,从 $\theta^{(0)}=0.5$ 开始,经过四次迭代,EM 算法收敛到观测后验分布的众数 0.6268.

另外,由(6.48)式不用迭代也可求出 θ 的估计. 事实上,(6.48)是一个迭代公式,若它收敛到 $\hat{\theta}$,则应有 $\hat{\theta}=(159\hat{\theta}+68)/(197\hat{\theta}+144)$,解之有 $\hat{\theta}=0.626\,821\,497$.

EM 算法的最大优点是简单和稳定. EM 算法的主要目的是提供一个简单的迭代算法来计算后验众数(或 MLE),人们自然会问,如此建立的 EM 算法能否达到预期要求,就是说,由 EM 算法得到的估计序列是否收敛,如果收敛其结果是否是 $p(\theta|Y)$ 的最大值或局部最大值. 为此,我们给出下述两个定理,记 EM 算法得到的估计序列为 $\theta^{(i)}$,$i=1,2,\cdots$,$L(\theta|Y)=\log p(\theta|Y)$.

定理 6.9　EM 算法在每一次迭代后均提高(观测)后验密度函数值,即

$$p(\theta^{(i+1)}|Y)\geqslant p(\theta^{(i)}|Y) \tag{6.49}$$

证明:由全概率公式

$$p(\theta,Z|Y)=p(Z|\theta,Y)p(\theta|Y)=p(\theta|Y,Z)p(Z|Y)$$

将上式后面两项取对数有

$$\log p(\theta|Y)=\log p(\theta|Y,Z)-\log p(Z|\theta,Y)+\log p(Z|Y) \tag{6.50}$$

设现有估计 $\theta^{(i)}$，将上式对 Z 关于 $p(Z|\theta^{(i)},Y)$ 求期望，有

$$\begin{aligned}\log p(\theta|Y) &= \int[\log p(\theta|Y,Z) - \log p(Z|\theta,Y) + \\ &\quad \log p(Z|Y)]p(Z|\theta^{(i)},Y)\mathrm{d}Z \\ &\triangleq Q(\theta|\theta^{(i)},Y) - H(\theta|\theta^{(i)},Y) + K(\theta^{(i)},Y) \end{aligned} \tag{6.51}$$

其中 $Q(\theta|\theta^{(i)},Y)$ 已在(6.44)中定义，

$$H(\theta|\theta^{(i)},Y) = \int\log[p(Z|\theta,Y)]p(Z|\theta^{(i)},Y)\mathrm{d}Z$$

$$K(\theta^{(i)},Y) = \int\log[p(Z|Y)]p(Z|\theta^{(i)},Y)\mathrm{d}Z$$

分别在(6.51)中取 θ 为 $\theta^{(i)}$ 和 $\theta^{(i+1)}$ 并相减，有

$$\begin{aligned}&\log p(\theta^{(i+1)}|Y) - \log p(\theta^{(i)}|Y) \\ &= [Q(\theta^{(i+1)}|\theta^{(i)},Y) - Q(\theta^{(i)}|\theta^{(i)},Y)] - \\ &\quad [H(\theta^{(i+1)}|\theta^{(i)},Y) - H(\theta^{(i)}|\theta^{(i)},Y)]\end{aligned}$$

由 Jensen 不等式，

$$\begin{aligned}&\mathrm{E}^{Z|\theta^{(i)},Y}\log\left(\frac{p(Z|\theta^{(i+1)},Y)}{p(Z|\theta^{(i)},Y)}\right) \\ &\leqslant \log\left\{\mathrm{E}^{Z|\theta^{(i)},Y}\left(\frac{p(Z|\theta^{(i+1)},Y)}{p(Z|\theta^{(i)},Y)}\right)\right\} = 0\end{aligned}$$

故 $H(\theta^{(i+1)}|\theta^{(i)},Y) - H(\theta^{(i)}|\theta^{(i)},Y) \leqslant 0$，而 $\theta^{(i+1)}$ 是使 $Q(\theta|\theta^{(i)},Y)$ 达最大的，显然

$$Q(\theta^{(i+1)}|\theta^{(i)},Y) - Q(\theta^{(i)}|\theta^{(i)},Y) \geqslant 0$$

证毕.

定理 6.10 (1) 如果 $p(\theta|Y)$ 有上界，则 $L(\theta^{(i)}|Y)$ 收敛到某个 L^*；

(2) 如果 $Q(\theta|\varphi)$ 关于 θ 和 φ 都连续，则在关于 L 的很一般的条件下，由 EM 算法得到的估计序列 $\theta^{(i)}$ 的收敛值 θ^* 是 L 的稳定点.

证明：(1)的证明由单调收敛定理立得；(2)的证明见文献[3].

关于定理 6.10，我们指出，定理的条件在大多数场合是满足的，定理的收敛性结论是针对(对数)后验密度函数值给出的. 而后验密度函数值序列的收敛性比估计序列本身的收敛性更有意义. Wu 也讨论了

估计序列本身的收敛性[13]，此处从略．另外，在定理 6.10 条件下，EM 算法的结果只能保证收敛到后验密度函数的稳定点，并不能保证收敛到极大值点，事实上，任何一种算法都很难保证其结果为极大值点．较为可行的办法是选取几个不同的初值进行迭代，然后在诸估计间加以选择，这可减轻初值选取对结果的影响．

例 6.14　表 6.3 的数据来自如下模型：在第 i 组中的第 j 个观测值为 y_{ij}，

$$(y_{ij}|\mu_i,\sigma) \sim N(\mu_i,\sigma^2),\quad j=1,\cdots,n_i;\quad i=1,\cdots,m$$

表 6.3　试验数据

组别 i	观测值
1	62,60,63,59
2	63,67,71,64,65,66
3	68,66,71,67,68,68
4	56,62,60,61,63,64,63,59

且诸 y_{ij} 独立．设 $\mu_i \sim N(\mu,\tau^2)$，记

$$\theta = (\mu,\log\sigma,\log\tau)$$

$$Y = \{y_{ij}, j=1,\cdots,n_i; i=1,\cdots,m\}$$

$$Z = (\mu_1,\cdots,\mu_m)$$

$$n = \sum_{i=1}^{m} n_i$$

θ 是我们感兴趣的未知参数，取其先验分布为

$$p(\mu,\log\sigma,\log\tau) \propto \tau$$

则由 Bayes 公式不难得到下述后验分布

$$p(\mu_1,\cdots,\mu_m,\mu,\log\sigma,\log\tau|Y)$$

$$\propto p(\mu,\log\sigma,\log\tau)\cdot$$

$$\prod_{i=1}^{m} p(\mu_i|\mu,\tau)\prod_{i=1}^{m}\prod_{j=1}^{n_i} p(y_{ij}|\mu_i,\sigma)$$

将之展开并取对数有

$$\log p(\mu_1,\cdots,\mu_m,\mu,\log\sigma,\log\tau|Y)$$
$$\propto -n\log\sigma-(m-1)\log\tau-\frac{1}{2\tau^2}\sum_{i=1}^{m}(\mu_i-\mu)^2-$$
$$\frac{1}{2\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\mu_i-y_{ij})^2$$

从而

$$\log p(\mu,\log\sigma,\log\tau|Y,Z)$$
$$\propto \log p(\mu_1,\cdots,\mu_m,\mu,\log\sigma,\log\tau|Y)$$
$$\propto -n\log\sigma-(m-1)\log\tau-\frac{1}{2\tau^2}\sum_{i=1}^{m}(\mu_i-\mu)^2-$$
$$\frac{1}{2\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\mu_i-y_{ij})^2 \tag{6.52}$$

直接极大化 $p(\mu,\log\sigma,\log\tau|Y)$ 求 μ,σ,τ 是很困难的，我们将 $Z=(\mu_1,\cdots,\mu_m)$ 看作潜在数据便可以采用 EM 算法来完成它的极大化.

在 E 步中，设已有 θ 的估计值 $\theta^{(t)}=(\mu^{(t)},\sigma^{(t)},\tau^{(t)})$，我们要在给定 Y 和 $\theta^{(t)}$ 的条件下求(6.52)式的期望. 因正态分布是 μ_i 的共轭先验分布，不难得到

$$(\mu_i|\theta^{(t)},Y)\sim N(\hat{\mu}_i^{(t)},V_i^{(t)})$$

其中

$$\hat{\mu}_i^{(t)}=\left(\frac{\mu}{(\tau^{(t)})^2}+\frac{n_i}{(\sigma^{(t)})^2}\bar{y}_{i.}\right)\Bigg/\left(\frac{1}{(\tau^{(t)})^2}+\frac{n_i}{(\sigma^{(t)})^2}\right)$$
$$V_i^{(t)}=\left(\frac{1}{(\tau^{(t)})^2}+\frac{n_i}{(\sigma^{(t)})^2}\right)^{-1}$$

这里 $\bar{y}_{i.}=\frac{1}{n_i}\sum_{j=1}^{n_i}y_{ij}$. 于是，对任意与 μ_i 无关的量 C，

$$\mathrm{E}[(\mu_i-C)^2|\theta^{(t)},Y]=[\mathrm{E}(\mu_i|\theta^{(t)},Y)-C]^2+\mathrm{Var}(\mu_i|\theta^{(t)},Y)$$
$$=(\hat{\mu}_i^{(t)}-C)^2+V_i^{(t)}$$

分别令 C 等于 μ 和 y_{ij} 即可得到(6.52)式在 Y 和 $\theta^{(t)}$ 给定下的期望(即 Q 函数)，

$$Q(\theta|\theta^{(t)},Y)=-n\log\sigma-(m-1)\log\tau-$$

$$\frac{1}{2\tau^2}\sum_{i=1}^{m}[(\hat{\mu}_i^{(t)}-\mu)^2+V_i^{(t)}]-$$

$$\frac{1}{2\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}[(y_{ij}-\hat{\mu}_i^{(t)})^2+V_i^{(t)}]-C_1 \qquad (6.53)$$

其中 C_1 是与 θ 无关的量.

在 M 步中,对(6.53)式的极大化是直接而简单的. 由 $Q(\theta|\theta^{(t)},Y)$ 对 μ,σ,τ 求导得

$$\mu^{(t+1)}=\frac{1}{m}\sum_{i=1}^{m}\hat{\mu}_i^{(t)}$$

$$\sigma^{(t+1)}=\left\{\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}[(y_{ij}-\hat{\mu}_i^{(t)})^2+V_i^{(t)}]\right\}^{1/2}$$

$$\tau^{(t+1)}=\left\{\frac{1}{m-1}\sum_{i=1}^{m}[(\hat{\mu}_i^{(t)}-\mu^{(t+1)})^2+V_i^{(t)}]\right\}^{1/2}$$

由此给出迭代公式.

该 EM 算法的实施是不难的. 取两组相近的初值,经过 8 次迭代,都收敛到 $\mu=64.0127,\sigma=2.4248,\tau=3.4993$,迭代中间结果见表6.4,误差控制取 $(\mu^{(t+1)}-\mu^{(t)})^2+(\sigma^{(t+1)}-\sigma^{(t)})^2+(\tau^{(t+1)}-\tau^{(t)})^2<1\times10^{-9}$.

表 6.4　EM 算法迭代结果

64.0000	64.0104	64.0124	64.0126	64.0127	64.0127	64.0127	64.0127
2.2900	2.4484	2.4198	2.4257	2.4246	2.4249	2.4248	2.4248
3.5600	3.5554	3.4981	3.5010	3.4993	3.4994	3.4994	3.4993
60.0000	63.4222	63.9585	64.0086	64.0124	64.0127	64.0127	64.0127
2.0000	2.6497	2.3954	2.4317	2.4237	2.4251	2.4248	2.4248
2.0000	3.4182	3.4120	3.4989	3.4969	3.4995	3.4993	3.4993

为了说明 EM 算法的收敛性,我们特地选择一组远离参数值的初值 $\mu=23,\sigma=5,\tau=1$ 进行 EM 计算,经过 82 次迭代后,它也收敛到上述结果,迭代过程见图 6.4.

EM 算法的简单可行性和稳定性使得 EM 算法得到广泛的应用,

特别是在指数型分布族中.

假设 $X=(Y,Z)$ 的分布为如下指数族,其密度函数为

$$p(x|\theta)=\exp\{\theta s(x)+c(\theta)-d(x)\}$$

取平坦先验分布 $\pi(\theta)\propto 1$,则由 Bayes 公式,

$$p(\theta|Y,Z)\propto\exp\{\theta s(x)+c(\theta)\}$$

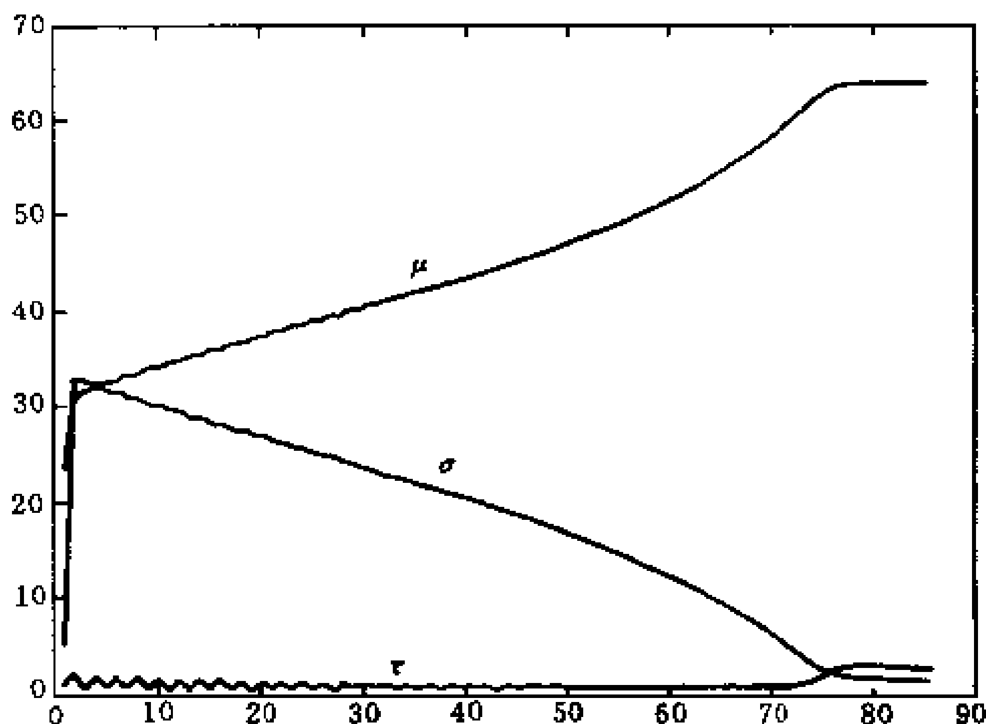


图 6.4　EM 算法迭代过程

在 E 步中,Q 函数可写出(可相差一个与 θ 无关的常数,下同)为

$$\begin{aligned}Q(\theta|\theta^{(i)},Y)&=\int\log[p(\theta|Y,Z)]p(Z|\theta^{(i)},Y)\mathrm{d}Z\\&=\theta\int s(x)p(Z|\theta^{(i)},Y)\mathrm{d}Z+c(\theta)\\&=\theta\mathrm{E}^{Z}[s(X)|\theta^{(i)},Y]+c(\theta)\end{aligned}$$

在 M 步中,由 $\dfrac{\partial Q(\theta|\theta^{(i)},Y)}{\partial\theta}=0$ 可给出

$$-\frac{\partial c(\theta)}{\partial\theta}=\mathrm{E}^{Z}[s(X)|\theta^{(i)},Y]$$

又由指数族性质(定理 1.14)知

$$\mathrm{E}[s(X)|\theta] \triangleq \mathrm{E}_\theta[s(X)] = -\frac{\partial \cdot(\theta)}{\partial\theta}$$

于是，EM 算法的迭代公式等价于解方程

$$\mathrm{E}[s(X)|\theta] = \mathrm{E}^Z[s(X)|\theta^{(i)}, Y]$$

这样求出的估计序列 $\theta^{(i)}$ 收敛到观测后验分布的众数.

§ 6.3.2　标准差

由前面介绍，EM 算法在一般条件下收敛到后验分布的众数，下面我们讨论 EM 估计的精度.

假设 EM 算法最后的结果是 $\hat{\theta}$，则 $\hat{\theta}$ 的渐近方差可用 Fisher 观测信息的倒数

$$V_0 = I_0^{-1} = \left(-\frac{\partial^2 \log p(\theta|Y)}{\partial\theta^2}\bigg|_{\hat{\theta}}\right)^{-1} \tag{6.54}$$

近似(参见 § 2.5.2). 因此，问题的关键是(6.54)的计算或近似计算，下面讨论两种算法.

§ 6.3.2.1　Louis 算法

由(6.50)式容易看出，

$$-\frac{\partial^2 \log p(\theta|Y)}{\partial\theta^2} = -\frac{\partial^2 \log p(\theta|Y,Z)}{\partial\theta^2} + \frac{\partial^2 \log p(Z|\theta,Y)}{\partial\theta^2}$$

将上式对 $p(Z|\theta,Y)$ 求期望，得到

$$\begin{aligned}
&-\frac{\partial^2 \log p(\theta|Y)}{\partial\theta^2} \\
&= -\int \frac{\partial^2 \log p(\theta|Y,Z)}{\partial\theta^2} p(Z|\theta,Y)\mathrm{d}Z + \\
&\quad \int \frac{\partial^2 \log p(Z|\theta,Y)}{\partial\theta^2} p(Z|\theta,Y)\mathrm{d}Z \\
&= \frac{-\partial^2 Q(\theta|\varphi,Y)}{\partial\theta^2}\bigg|_{\varphi=\theta} - \\
&\quad \frac{-\partial^2 H(\theta|\varphi,Y)}{\partial\theta^2}\bigg|_{\varphi=\theta}
\end{aligned}$$

其中 Q 和 H 如(6.51)中定义,$-\frac{\partial^2 Q}{\partial\theta^2}$和$-\frac{\partial^2 H}{\partial\theta^2}$分别称为完全信息和缺损信息.于是我们得到所谓的"缺损信息原则",即

观测信息 = 完全信息 − 缺损信息

Louis 曾得到了一个重要结论[7]:

$$-\int \frac{\partial^2 \log p(Z|\theta,Y)}{\partial\theta^2} p(Z|\theta,Y)\mathrm{d}Z = \mathrm{Var}\left(\frac{\partial \log p(\theta|Y,Z)}{\partial\theta}\right)$$

故

$$\begin{aligned}&-\frac{\partial^2 \log p(\theta|Y)}{\partial\theta^2}\\ &= -\int \frac{\partial^2 \log p(\theta|Y,Z)}{\partial\theta^2} p(Z|\theta,Y)\mathrm{d}Z - \\ &\quad \mathrm{Var}\left(\frac{\partial \log p(\theta|Y,Z)}{\partial\theta}\right)\end{aligned} \tag{6.55}$$

其中涉及到的方差表达式应对 Z 关于(θ,Y)的条件分布 $p(Z|\theta,Y)$进行,将(6.55)式中的 θ 换为 $\hat{\theta}$,其倒数即为 $\hat{\theta}$ 的渐近方差.

例 6.15 在例 6.13 中,

$$\frac{\partial \log p(\theta|Y,Z)}{\partial\theta} \propto \frac{y_1 - Z + y_4}{\theta} - \frac{y_2 + y_3}{1-\theta}$$

记最后得到的估计为 $\hat{\theta} \triangleq 0.6268$,则

$$\begin{aligned}&-\int \frac{\partial^2 \log p(\hat{\theta}|Y,Z)}{\partial\theta^2} p(Z|\hat{\theta},Y)\mathrm{d}Z\\ &= \int\left[\frac{y_1 - Z + y_4}{\hat{\theta}^2} + \frac{y_2 + y_3}{(1-\hat{\theta})^2}\right] p(Z|\hat{\theta},Y)\mathrm{d}Z\\ &= \frac{y_1 + y_4 - \mathrm{E}(Z|\hat{\theta},Y)}{\hat{\theta}^2} + \frac{y_2 + y_3}{(1-\hat{\theta})^2}\\ &= \frac{159 - 125 \cdot 2/(2+0.6268)}{0.6268^2} + \frac{38}{(1-0.6268)^2}\\ &= 435.3\end{aligned}$$

而

$$\mathrm{Var}\left(\left.\frac{\partial \log p(\theta|Y,Z)}{\partial\theta}\right|_{\hat{\theta}}\right) = \frac{\mathrm{Var}(Z|\hat{\theta},Y)}{\hat{\theta}^2}$$

$$= 125 \cdot \left(\frac{2}{2+\hat{\theta}}\right) \cdot \left(\frac{\hat{\theta}}{2+\hat{\theta}}\right) / \hat{\theta}^2$$
$$= 57.8$$

于是

$$-\left.\frac{\partial^2 \log p(\theta|Y)}{\partial \theta^2}\right|_{\hat{\theta}} = 435.3 - 57.8 = 377.5$$

$\hat{\theta}$ 的渐近标准差为 $377.5^{-1/2} = 0.05$.

§6.3.2.2 模拟计算

例 6.15 给出了一个能够显式计算的例，在许多场合，给出(6.55)的显式表达式是很困难的，甚至是不可能的. 此时可考虑采用模拟的办法给出(6.55)式右端的近似.

设我们可从分布 $p(Z|\hat{\theta},Y)$ 抽取 m 个样本，记为 $z_1,\cdots,z_m$，如果 m 充分大，则(6.55)式可近似如下

$$-\int \left.\frac{\partial^2 \log p(\theta|Y,Z)}{\partial \theta^2}\right|_{\hat{\theta}} p(Z|\hat{\theta},Y)\mathrm{d}Z$$
$$\approx -\frac{1}{m}\sum_{j=1}^{m} \left.\frac{\partial^2 \log p(\theta|Y,z_j)}{\partial \theta^2}\right|_{\hat{\theta}}$$
$$\mathrm{Var}\left(\left.\frac{\partial \log p(\theta|Y,Z)}{\partial \theta}\right|_{\hat{\theta}}\right)$$
$$\approx \frac{1}{m}\sum_{j=1}^{m}\left(\left.\frac{\partial \log p(\theta|Y,z_j)}{\partial \theta}\right|_{\hat{\theta}}\right)^2 -$$
$$\left[\frac{1}{m}\sum_{j=1}^{m} \left.\frac{\partial \log p(\theta|Y,z_j)}{\partial \theta}\right|_{\hat{\theta}}\right]^2$$

例 6.16 在例 6.13 中，$p(Z|Y,\theta)$ 为二项分布，在 EM 算法收敛到 $\hat{\theta}=0.6268$ 后，$p(Z|\hat{\theta},Y)$ 为 $b(y_1, 2/(2+\hat{\theta})) = b(125, 0.7614)$，我们可由二项分布 $b(125, 0.7614)$ 抽取 m 个独立观测值，然后近似计算(6.55)式.

由(6.47)，有

$$\left.\frac{\partial \log p(\theta|Y,z)}{\partial \theta}\right|_{\hat{\theta}} \propto \frac{y_1 - z + y_4}{\hat{\theta}} - \frac{y_2 + y_3}{1-\hat{\theta}}$$

$$= \frac{125+34-z}{0.6268} - \frac{18+25}{0.3732}$$

$$= 154.8474 - 1.5954z$$

$$-\left.\frac{\partial^2 \log p(\theta|Y,z)}{\partial\theta^2}\right|_{\hat\theta} \propto \frac{y_1 - z + y_4}{\hat\theta^2} + \frac{y_2 + y_3}{(1-\hat\theta)^2}$$

$$= \frac{125+34-z}{0.6268^2} + \frac{18+20}{0.3732^2}$$

$$= 677.5407 - 2.5453z$$

在得到随机数 $z_1,\cdots,z_m$ 后，记 $\bar{z}=\frac{1}{m}\sum_{j=1}^{m} z_j$, $\overline{z^2}=\frac{1}{m}\sum_{j=1}^{m} z_j^2$，有

$$-\int \left.\frac{\partial^2 \log p(\theta|Y,Z)}{\partial\theta^2}\right|_{\hat\theta} p(Z|\hat\theta,Y)\mathrm{d}Z \approx -\frac{1}{m}\sum_{j=1}^{m}(677.5407 - 2.5453z_j)$$

$$= 677.5407 - 2.5453\bar{z}$$

$$\mathrm{Var}\left(\left.\frac{\partial \log p(\theta|Y,Z)}{\partial\theta}\right|_{\hat\theta}\right) \approx \frac{1}{m}\sum_{j=1}^{m}(154.8474 - 1.5954z_j)^2 - (154.8474 - 1.5954\bar{z})^2$$

$$= 2.5453[\overline{z^2} - (\bar{z})^2]$$

至此可给出(6.55)的近似值为 $677.5407-2.5453(\bar{z}+\overline{z^2}-(\bar{z})^2)$，我们在计算机上抽取 10 000 个 $b(125,0.7614)$ 的随机数，近似计算的结果为 57.4563，而精确值为 57.8.

倘 $p(Z|Y,\theta)$ 不能直接抽样，可考虑采用§6.1 节中介绍的各种抽样方法.

§6.3.3 GEM 算法

EM 算法得到广泛应用的一个重要原因是在 M 步中，求极大化的方法与完全数据下求极大化的方法完全一样. 在许多场合，这样的极大化有显式表示，然而并不总是这样，有时要找使(6.45)式右端达最大的 θ 是困难的，一个较简单的方法是找一个 $\theta^{(i+1)}$，使得

$$Q(\theta^{(i+1)}|\theta^{(i)},Y) > Q(\theta^{(i)}|\theta^{(i)},Y) \tag{6.56}$$

由(6.44)和(6.56)组成的 EM 算法称为广义 EM 算法(GEM 算法).

从定理 6.9 的证明可以看出,GEM 算法也保证有 $p(\theta^{(i-1)}|Y)\geqslant p(\theta^{(i)}|Y)$,进一步,略微改变一下条件,定理 6.10 对 GEM 算法也是成立的[13].

当 θ 是一维参数时,现已有许多极大化方法可以使用,因而没有必要考虑 GEM 算法. 对多维参数 $\theta=(\theta_1,\cdots,\theta_k)$,Meng 和 Rubin 提出一种特殊的 GEM 算法[14],他们称之为 ECM(Expectation/Conditional maximization)算法,该算法保留了 EM 算法的简单性和稳定性,它将原先 EM 算法中的 M 步(极大化)分解为如下 k 次条件极大化:在第 $i+1$ 次迭代中,记 $\theta^{(i)}=(\theta_1^{(i)},\cdots,\theta_k^{(i)})$,在得到 $Q(\theta|\theta^{(i)},Y)$后,首先在 $\theta_2^{(i)},\cdots,\theta_k^{(i)}$ 保持不变的条件下求使 $Q(\theta|\theta^{(i)},Y)$达到最大的 $\theta_1^{(i+1)}$,然后在 $\theta_1=\theta_1^{(i+1)},\theta_j=\theta_j^{(i)},j=3,\cdots,k$ 的条件下求使 $Q(\theta|\theta^{(i)},Y)$达到最大的 $\theta_2^{(i+1)}$,如此继续,经过 k 次条件极大化,我们得到了一个 $\theta^{(i+1)}$,完成了一次迭代. 该 ECM 算法很简单,它是一种 GEM 算法,满足 GEM 算法的所有性质,譬如收敛性.

§ 6.3.4 Monte Carlo EM 算法

EM 算法由求期望(E 步)和求极值(M 步)两部分组成,M 步由于等同于完全数据的处理,通常比较简便,而在 E 步中,有时要获得期望的显式表示是不可能的,即使近似计算也很困难,这时可用 Monte Carlo 方法来完成,这就是所谓的 Monte Carlo EM(MCEM)方法,它将 E 步改为:

(E1) 由 $p(Z|\theta^{(i)},Y)$随机地抽取 m 个随机数 $z_1,\cdots,z_m$;

(E2) 计算 $\hat{Q}(\theta|\theta^{(i)},Y)=\frac{1}{m}\sum_{j=1}^{m}\log p(\theta|z_j,Y)$.

由大数定律,只要 m 足够大,$\hat{Q}(\theta|\theta^{(i)},Y)$与 $Q(\theta|\theta^{(i)},Y)$很接近,从而我们可以在 M 步中对 $\hat{Q}(\theta|\theta^{(i)},Y)$求极大化.

在 MCEM 算法中有二点是主要要考虑的. 一是 m 大小的确定,从精度角度来讲,m 自然越大越好,但过大的 m 使得计算的效太低,一般在开始时 m 不需很大. 另一点是收敛性的判断,因在 E 步中是采用的 Monte Carlo 方法,若要求这样得到的 $\theta^{(i)}$ 收敛到一点显然是不现实

的. 在 MCEM 中,收敛性的判别往往可借助图形来进行. 若经过若干次迭代后,迭代值围绕直线 $\theta=\theta^*$ 小幅波动,则可以认为算法收敛了. 此时,为增加估计精度,可增加 m 的值再运行一段时间,就可停止.

MCEM 得到的估计的标准差同样可由 §6.3.2 节中的方法加以估计,只需将那里的 Q 改为 $\hat{Q}$ 即可,此处不再讨论.

例 6.17 再考虑例 6.13. 假设在第 i 次迭代开始时已有 θ 的估计 $\theta^{(i)}$,因 $(Z|\theta^{(i)},Y)\sim b(y_1,p_i)$, $p_i=\theta^{(i)}/(2+\theta^{(i)})$. 因此,我们可从 $b(y_1,p_i)$ 抽 m 个随机数 $z_1,\cdots,z_m$,并计算 Q 函数

$$\hat{Q}(\theta|\theta^{(i)},Y)=\frac{1}{m}\sum_{j=1}^{m}\log p(\theta|z_j,Y)$$
$$=(\bar{z}+y_4)\log\theta+(y_2+y_3)\log(1-\theta)$$

在 M 步中,将上式对 θ 求导并令其为 0,易有

$$\theta^{(i+1)}=\frac{\bar{z}+y_4}{\bar{z}+y_4+y_2+y_3}$$

将之在计算机上进行运算,取 θ 的初值为 0.4,开始时取 $m=20$,在经过 10 次迭代后可看出,迭代得到的 $\theta^{(i)}$ 值呈小幅波动情形. 将 m 增加到 200,再运行 10 次,结果见图 6.5,可用后 10 次的均值 0.6271 作为 θ 的估计.

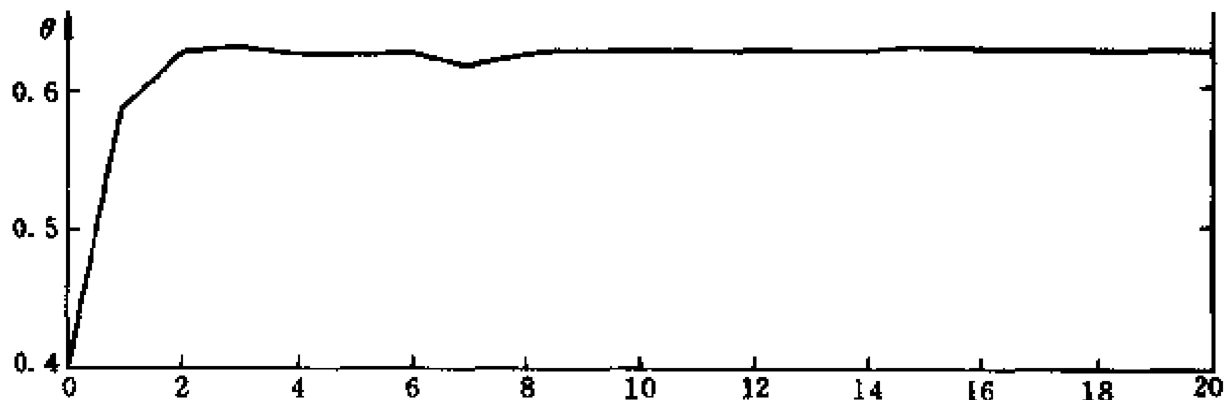


图 6.5 运行中的迭代结果

§ 6.4　Markov Chain Monte Carlo(MCMC)方法

EM 算法得到的是后验分布的众数，有时我们希望得到其它一些后验量，比如后验均值、后验方差、后验分布的分位数，等等．计算这些后验量都可归结为关于后验分布的积分计算．具体地，设 $\pi(x)$，$x\in\mathscr{X}$ 为后验分布，我们要计算的后验量可写成某函数 $f(x)$ 关于 $\pi(x)$ 的期望

$$E_\pi f = \int_{\mathscr{X}} f(x)\pi(x)dx \tag{6.57}$$

对于较简单的后验分布，我们可以直接计算(6.57)或利用正态近似、数值积分、静态 Monte Carlo 等近似计算方法．但当后验分布很复杂时，这些方法都难以实施，而在实际中，观测后验分布往往是复杂的、高维的、非标准形式的分布．因此，我们必须探讨一些新的计算方法．Markov Chain Monte Carlo(MCMC)就是最近发展起来的一种简单且行之有效的 Bayes 计算方法．MCMC 方法在统计物理学中得到广泛应用已有四十多年的历史，但它在 Bayes 统计、显著性检验、极大似然估计等方面的应用则是近十年内的事情．

§ 6.4.1　基本思路

MCMC 方法的基本思想是通过建立一个平稳分布为 $\pi(x)$ 的 Markov 链来得到 $\pi(x)$ 的样本，基于这些样本就可以作各种统计推断．比如，若得到了 $\pi(x)$ 的样本 $X^{(1)},\cdots,X^{(n)}$，则(6.57)可估计为

$$\hat{f}_n = \frac{1}{n}\sum_{i=1}^{n} f(x^{(i)}) \tag{6.58}$$

这便是 Monte Carlo 积分．我们知道，当 $X^{(1)},\cdots,X^{(n)}$ 独立时，由大数定律有

$$\hat{f}_n \xrightarrow{a.s.} E_\pi f,\quad n\to\infty \tag{6.59}$$

但当 $X^{(1)},\cdots,X^{(n)}$ 是平稳分布为 $\pi(x)$ 的 Markov 过程的样本时，(6.59)式也成立．

以上是 MCMC 方法的简单叙述. 为进一步说明 MCMC 方法的做法,我们先简单回顾一下 Markov 链. 假定我们用如下方法产生一个随机变量序列$\{X^{(0)},X^{(1)},X^{(2)},\cdots\}$:在任一时刻$t(t\geqslant 0)$,序列中下一时刻$t+1$处的$X^{(t+1)}$由条件分布$P(x|X^{(t)})$产生,它只依赖于时刻$t$处的当前状态而与时刻$t$以前的历史状态$\{X^{(0)},X^{(1)},\cdots,X^{(t-1)}\}$无关,这样一个随机变量序列称为 Markov 链.

一个问题是$X^{(0)}$对$X^{(t)}$有什么影响,亦即是说,在给定$X^{(0)}$而没有$\{X^{(1)},\cdots,X^{(t-1)}\}$的信息情况下,$X^{(t)}$的分布情况怎样? 我们将$X^{(0)}$给定下$X^{(t)}$的条件分布记为$P^{(t)}(X^{(t)}|X^{(0)})$,那么,不同的$X^{(0)}$的取值对$X^{(t)}$的分布是否有显著影响?我们看一个例子,考虑从$N(0.5X^{(t-1)},1)$抽取$X^{(t)}$,对不同的初始值$X^{(0)}$,我们可以得到不同的 Markov 链的实现,但当t充分大后,$X^{(t)}$的分布已经与$X^{(0)}$无关(见图 6.6). 这表明,不管初始值取什么,$X^{(t)}$的分布收敛到同一个分布,即所谓的平稳分布.

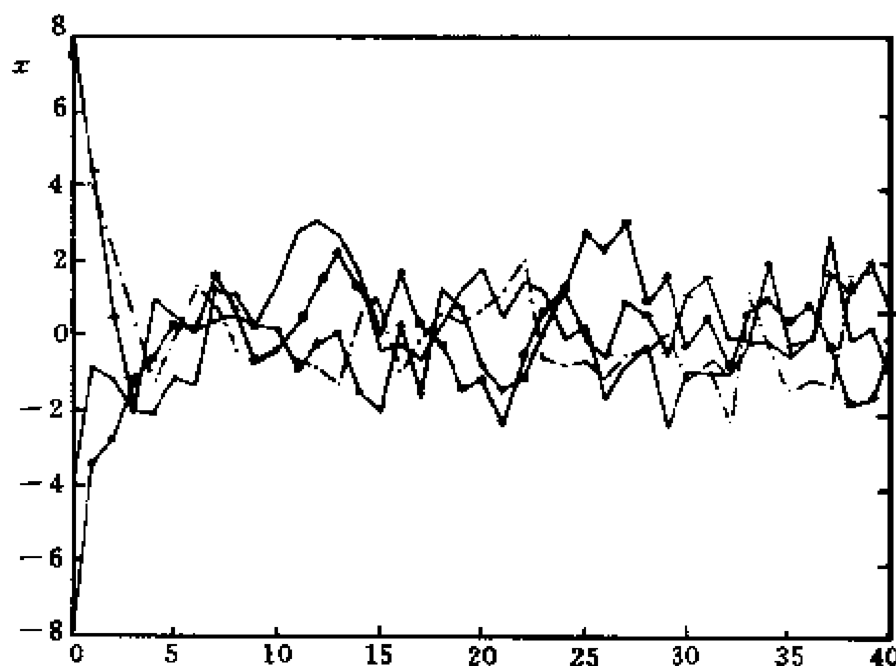


图 6.6 Markov 链的实现值

一般地,令$\{X^{(t)}\}_{t\geqslant 0}$为$\mathscr{X}$上的 Markov 链,其一步转移概率函数为

$$p(x,x')\triangleq P(x\to x')=P(X^{(t+1)}=x'|X^{(t)}=x)(\text{离散})$$

或

$$P(x \to B) = \int_B p(x, x')\mathrm{d}x' \text{（连续）}$$

$p(\cdot,\cdot)$称为该Markov链的转移核. 常假定$p(\cdot,\cdot)$与t无关，即该Markov链是时间齐次的. t步转移概率函数为

$$p(t; x, x') \triangleq P(X^{(t+s)} = x' \mid X^{(s)} = x)$$

记$X^{(0)}$的分布为$\mu(x) = P(X^{(0)} = x)$，则经过t步后$X^{(t)}$的边际分布记为

$$\mu^t(x) = P(X^{(t)} = x)$$

如果$\pi(x)$满足

$$\int p(x, x')\pi(x)\mathrm{d}x = \pi(x'), \quad \forall\, x' \in \mathscr{X}$$

则称$\pi(x)$为转移核$p(\cdot,\cdot)$的平稳分布. 比如，在上面的例子中，理论上可算出平稳分布为$N(0.0, 1.33)$，模拟结果与之吻合.

作为起始状态，$X^{(0)}$最好具有分布$\pi(x)$，于是，由平稳分布的定义保证任一$X^{(t)}$的边际分布也是$\pi(x)$，然而这在需要应用MCMC时通常做不到. 正因为从$\pi(x)$难以直接取样我们才借助MCMC方法. 事实上，我们并不需要起始状态的边际分布就是$\pi(x)$，从图6.6可以看出，从不同的$X^{(0)}$出发，链经过一段时间的迭代后，可以认为各个时刻的$X^{(t)}$的边际分布都是平稳分布$\pi(x)$，此时，我们称它收敛了. 而在收敛出现以前的一段时间，比如m次迭代中，各状态的边际分布还不能认为是$\pi(x)$，因此，我们在使用(6.59)估计$\mathrm{E}_\pi f$时，应把前面的m个迭代值去掉，而用后面的$n-m$个迭代结果来估计，即

$$\hat{f}_{mn} = \frac{1}{n-m}\sum_{t=m+1}^{n} f(X^{(t)}) \tag{6.60}$$

上式称为遍历平均. 由众知的遍历性定理，有$\hat{f}_{mn} \to \mathrm{E}_\pi f$，$(n \to \infty)$.

从模拟计算的角度看，我们构造的转移核使已知的概率分布$\pi(x)$为平稳分布. 因此，在采用MCMC方法时，转移核的构造具有至关重要的作用. 不同的MCMC方法，如Gibbs抽样方法、Metropolis方法等，往往也就是转移核的构造方法不同，这将在后面具体介绍.

至此，我们可以把MCMC方法概括为如下三步：

(1) 在 $\mathscr{X}$ 上选一个"合适"的 Markov 链,使其转移核为 $p(\cdot,\cdot)$,"这里合适"的含义主要指 $\pi(x)$ 应是其相应的平稳分布;

(2) 由 $\mathscr{X}$ 中某一点 $X^{(0)}$ 出发,用(1)中的 Markov 链产生点序列 $X_1,\cdots,X_n$;

(3) 对某个 m 和大的 n,任一函数 $f(x)$ 的期望估计如下

$$\hat{\mathrm{E}}_\pi f = \frac{1}{n-m}\sum_{t=m+1}^{n} f(X^{(t)})$$

MCMC 有许多研究专题,比如 Markov 链的收敛性判断(m 大小的确定),链的长度(n 的大小)的确定,估计误差等等.下面我们主要讨论转移核的构造.

§6.4.2 满条件分布

MCMC 主要应用在多变量,非标准形式,且各变量间相互不独立时分布的模拟.显然,在作如此模拟时,条件分布起很大作用.

令 $\boldsymbol{x}=(x_1,\cdots,x_n)'$,我们总可以写出

$$\pi(\boldsymbol{x}) = \prod_{i=1}^{n}\pi(x_i|x_{<i}) \tag{6.61}$$

其中 $x_{<i}=\{x_j,j<i\}$.如果(6.61)式中右端的各个因子(可将 $x_1,\cdots,x_n$ 重新排列)能够直接模拟,则只需要进行静态模拟(模拟过程中不改变抽样分布)即可,而不需要应用 MCMC 方法.实际中,很难做到使(6.61)式中右端的各因子能直接模拟,因此需进行动态模拟(抽样分布随模拟的进行而改变),如 MCMC,此时满条件分布扮演了一个重要角色.

MCMC 方法大多建立在形如 $\pi(x_T|x_{-T})$ 的条件分布上,其中 $x_T=\{x_i,i\in T\}$,$x_{-T}=\{x_i,i\notin T\}$,$T\subset\mathbf{N}=\{1,\cdots,n\}$.注意到,在上述条件分布中,所有的变量全部出现了(或出现在条件中,或出现在变元中),这种条件分布称为满条件的(Full conditionals).

在导出满条件分布时,应注意到这样一个简单而有效的事实:对任意的 $x\in\mathscr{X}$ 和任意的 $T\subset\mathbf{N}$,

$$\pi(x_T|x_{-T}) = \frac{\pi(x)}{\int \pi(x)\mathrm{d}x_T} \propto \pi(x) \tag{6.62}$$

亦即，在 $\pi(x)$ 的乘积项中，只有与 x_T 有关的项需保留. 等价地，若 $x, x' \in \mathscr{X}$，且 $x_{-T} = x'_{-T}$，则

$$\frac{\pi(x'_T|x'_{-T})}{\pi(x_T|x_{-T})} = \frac{\pi(x')}{\pi(x)} \tag{6.63}$$

(6.62)和(6.63)很重要，因为后验分布密度函数通常是一些乘积项. 同时，复杂的后验分布的正则化常数往往无法计算，而 MCMC 方法的一个显著优点是，在应用 MCMC 时，$\pi(x)$ 以及满条件分布可以相差一个比例常数.

一般地，用 y 表示观测数据，$\boldsymbol{x} = (\theta, \varphi, z)$，其中，$\theta, \varphi, z$ 分别表示参数，超参数和缺损数据，则上述 $\pi(\boldsymbol{x})$ 可写为 $\pi(\boldsymbol{x}|y)$.

$$\pi(\boldsymbol{x}|y) \propto p(y,z|\theta)\pi(\theta|\varphi)\pi(\varphi)$$

其中，$p(y,z|\theta)$ 表示完全数据的密度函数，$\pi(\theta|\varphi)$ 表示先验分布，$\pi(\varphi)$ 为超参数的分布. 由(6.62)，各变量的满条件分布可如下给出：

$$\pi(\theta_i|\theta_{-i},\varphi,z,y) \propto p(y,z|\theta)\pi(\theta_i|\theta_{-i},\varphi)$$

$$\pi(\varphi_j|\theta,\varphi_{-j},z,y) \propto \pi(\theta|\varphi)\pi(\varphi)$$

$$\pi(z_k|\theta,\varphi,z_{-k},y) \propto p(y,z|\theta)$$

其中 $\theta_{-i} = \{\theta_j : j \neq i\}$，$\varphi_{-j}, z_{-k}$ 类似地定义.

在实用中，上述满条件分布有时可能会更简化.

例 6.18　设 X_1, X_2 的联合密度函数为

$$\pi(x_1,x_2) \propto \exp\left\{-\frac{1}{2}(x_1-1)^2(x_2-1)^2\right\}$$

则我们可容易地给出其满条件分布为

$$\begin{aligned}
\pi(x_1|x_2) &\propto \pi(x_1,x_2) \\
&\propto \exp\left\{-\frac{1}{2}(x_1-1)^2(x_2-1)^2\right\} \\
&= N(1,(x_2-1)^{-2})
\end{aligned}$$

$$\begin{aligned}
\pi(x_2|x_1) &\propto \pi(x_1,x_2) \\
&\propto \exp\left\{-\frac{1}{2}(x_2-1)^2(x_1-1)^2\right\}
\end{aligned}$$

$$= N(1,(x_1-1)^{-2})$$

例 6.19 设 $y_1,\cdots,y_n$ 是来自 $N(\mu,\tau^{-1})$ 的样本，并设 (μ,τ) 的先验分布分别为 $\mu\sim N(0,1)$，$\tau\sim Ga(2,1)$，且 μ 与 τ 独立．记 $Y=(y_1,\cdots,y_n)$，则 Y,μ,τ 的联合密度函数为

$$p(Y,\mu,\tau)=(2\pi)^{-\frac{n+1}{2}}\tau^{\frac{n}{2}}\exp\left\{-\frac{\tau}{2}\sum(y_i-\mu)^2-\frac{\mu^2}{2}\right\}\tau\mathrm{e}^{-\tau}$$

其后验分布为

$$\pi(\mu,\tau|Y)=\frac{p(Y,\mu,\tau)}{\int p(Y,\mu,\tau)\mathrm{d}\mu\mathrm{d}\tau}$$

满条件分布为

$$\begin{aligned}\pi(\mu|\tau,Y)&\propto\exp\left\{-\frac{\tau}{2}\sum(y_i-\mu)^2-\frac{\mu^2}{2}\right\}\\&\propto\exp\left\{-\frac{1}{2}(1+n\tau)\left[\mu-\frac{\tau\sum y_i}{1+n\tau}\right]^2\right\}\\&=N\left(\frac{\tau\sum y_i}{1+n\tau},(1+n\tau)^{-1}\right)\end{aligned}$$

$$\begin{aligned}\pi(\tau|\mu,Y)&\propto\tau^{n/2}\exp\left\{-\frac{\tau}{2}\sum(y_i-\mu)^2\right\}\tau\mathrm{e}^{-\tau}\\&=\tau^{1+\frac{n}{2}}\exp\left\{-\tau\left[1+\frac{1}{2}\sum(y_i-\mu)^2\right]\right\}\\&=Ga\left(2+\frac{n}{2},1+\frac{1}{2}\sum(y_i-\mu)^2\right)\end{aligned}$$

满条件分布并不都能表示为显式形式，看下例．

例 6.20 考虑 Logistic 回归模型，假设 $y_i\sim b(1,p_i)$，且诸 y_i 独立，$i=1,\cdots,n$，其中 $p_i=\{1+\mathrm{e}^{-(\alpha+\beta x_i)}\}^{-1}$，并假设 α,β 的先验分布分别为 $\alpha\sim N(0,1)$，$\beta\sim N(0,1)$，α 与 β 独立，则我们可写出各满条件分布，譬如，β 的满条件分布为

$$\pi(\beta|\alpha,Y)\propto\mathrm{e}^{-\frac{1}{2}\beta^2}\prod_{i=1}^{n}\{[\mathrm{e}^{-(\alpha+\beta x_i)}]^{1-y_i}[1+\mathrm{e}^{-(\alpha+\beta x_i)}]^{-1}\}$$

这个满条件分布是无法进一步简化的．幸运的是，虽然上述满条件分布很复杂，我们仍然可以运用一些抽样方法来抽取它的样本，比如 § 6.1

介绍的各种方法,以及其它方法.

§6.4.3 Gibbs 抽样

最简单、应用最广泛的 MCMC 方法是 Gibbs 抽样,它是由 Geman 和 Geman 最初[15]命名提出的,它的想法很直观.

设 $X=(X_1,\cdots,X_n)$ 的密度函数为 $\pi(x)$,任意固定 $T\subset\mathbf{N}$,在给定 $X_{-T}=x_{-T}$ 条件下,如下定义随机变量 $X'=(X'_1,\cdots,X'_n)$: $X'_{-T}=X_{-T}$而 X'_T 具有密度函数 $\pi(x'_T|x_{-T})$,则对任一可测集 B,

$$
\begin{aligned}
P(X'\in B)&=\int_B\pi(x'_{-T})\pi(x'_T|x'_{-T})\mathrm{d}x'\\
&=\int_B\pi(x')\mathrm{d}x'\\
&=\pi(B)
\end{aligned}
$$

因而 X' 的密度函数也是 $\pi(x)$.

上述过程定义了一个由 X 到 X'的转移核,且其相应的平稳分布是 π. 这样构造的 MCMC 称为 Gibbs 抽样(Gibbs sampler). 当 T 只含有一个元素时称为单元素 Gibbs 抽样(single-site Gibbs sampler),单元素 Gibbs 抽样是最简单的 MCMC. 比如 $T=\{i\}$时、单元素 Gibbs 抽样是在给定 $x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_n$ 下,由 x_i 关于 $(x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_n)$的满条件分布抽样,因为它只涉及单变量抽样,这使之最具吸引力.

在 Gibbs 抽样的构造之初,我们假设 X 具有密度函数 $\pi(x)$,这在实际中往往做不到,但这并不影响 Gibbs 抽样的实施. 应用中,我们可对 $i=1,\cdots,n$ 重复使用 Gibbs 抽样,在一般的条件下,这样的迭代依分布收敛到 π. 下面我们将单元素 Gibbs 抽样具体化写出.

在给出起始点 $x^{(0)}=(x_1^{(0)},\cdots,x_n^{(0)})$后,假定第 t 次迭代开始时的估计值为 $x^{(t-1)}$,则第 t 次迭代分为如下 n 步:

(1) 由满条件分布 $\pi(x_1|x_2^{(t-1)},\cdots,x_n^{(t-1)})$抽取 $x_1^{(t)}$;

……

(i) 由满条件分布 $\pi(x_i|x_1^{(t)},\cdots,x_{i-1}^{(t)},x_{i+1}^{(t-1)},\cdots,x_n^{(t-1)})$抽取 $x_i^{(t)}$;

……

(n) 由满条件分布 $\pi(x_n|x_1^{(t)},\cdots,x_{n-1}^{(t)})$ 抽取 $x_n^{(t)}$.

记 $x^{(t)}=(x_1^{(t)},\cdots,x_n^{(t)})$，则 $x^{(1)},x^{(2)},\cdots,x^{(t)},\cdots$ 是 Markov 链的实现值，其由 x 至 x' 的转移概率函数为

$$p(x,x')=\pi(x_1|x_2,\cdots,x_n)\pi(x_2|x_1',x_3,\cdots,x_n)\cdots\pi(x_n|x'_1,\cdots,x'_{n-1}).$$

由前述讨论知 $\pi(x)$ 为其平稳分布.

要使 Gibbs 抽样能够应用到实际问题中去，我们还必须知道什么时候 Gibbs 抽样可以停止(即收敛)，以及怎样对 Gibbs 抽样得到的数据进行分析. 这是一个重要而困难的问题，下面我们只作一些简单的介绍.

关于 Gibbs 抽样的收敛性判断，几乎没有简单而有效的方法，在实用中，通常可采取两种办法来进行判断.

方法之一是用 Gibbs 抽样同时产生多个 Markov 链，在经过一段时间后，如果这几条链稳定下来，则 Gibbs 抽样收敛了. 图 6.7 是这个方法的一个直观说明，它用 Gibbs 抽样同时产生 9 条 Markov 链，并把其中的一个参数 θ 的实现值作成散点图，由图可清楚地看到，在经过约 4400 次迭代后，Gibbs 抽样收敛了.

另一个判断 Gibbs 抽样收敛的方法是看遍历均值是否已经收敛，比如，我们在由 Gibbs 抽样得到的链中每隔一段距离计算一次参数的遍历均值，为使用来计算平均值的变量近似独立，通常可每隔一段取一个样本，当这样算得的均值稳定后，可认为 Gibbs 抽样收敛. 下面我们用两个例来加以说明.

例 6.21　Gelfand 和 Smith[16],[17] 考虑如下的例. 设某试验可能有五个结果，其出现的概率分别为

$$\frac{\theta}{4}+\frac{1}{8},\quad \frac{\theta}{4},\quad \frac{\eta}{4},\quad \frac{\eta}{4}+\frac{3}{8},\quad \frac{1}{2}(1-\theta-\eta)$$

它含有两个未知参数 θ 和 η，且都介于 0 与 1 之间. 现有 22 次试验的结果观测值为 $y=(y_1,\cdots,y_5)=(14,1,1,1,5)$，$y_i$ 表示 22 次试验中出现第 i 个结果的次数，$i=1,\cdots,5$. Gelfand 和 Smith 将 y_1,y_4 各分解成两

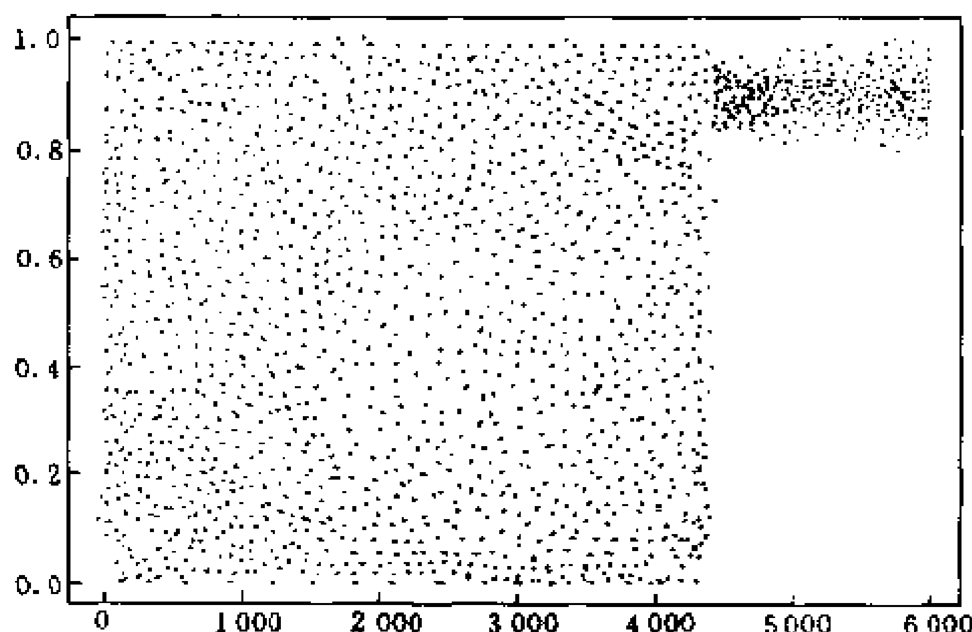


图 6.7　Gibbs 抽样迭代过程

部分，即令 $Y_1=Z_1+(Y_1-Z_1)$，$Y_4=Z_2+(Y_4-Z_2)$ 使

$$(Z_1,Y_1-Z_1,Y_2,Y_3,Z_2,Y_4-Z_2,Y_5)$$
$$\sim M\left[22;\frac{\theta}{4},\frac{1}{8},\frac{\theta}{4},\frac{\eta}{4},\frac{\eta}{4},\frac{3}{8},\frac{1}{2}(1-\theta-\eta)\right]$$

其中 M 表示多项分布，Z_1,Z_2 是不可观测的，可看作缺损数据. 取平坦分布作为 (θ,η) 的先验分布，即

$$\pi(\theta,\eta)\propto 1$$

若令 $z=(z_1,z_2)$ 为 $Z=(Z_1,Z_2)$ 的取值，则 (y,z) 的似然函数和后验分布分别为

$$L(y,z|\theta,\eta)\propto\left(\frac{\theta}{4}\right)^{z_1+y_2}\left(\frac{1}{8}\right)^{y_1-z_1}\times$$
$$\left(\frac{\eta}{4}\right)^{y_3+z_2}\left(\frac{3}{8}\right)^{y_4-z_2}\left(\frac{1-\theta-\eta}{2}\right)^{y_5}$$
$$\pi(\theta,\eta|y,z)\propto\theta^{z_1+y_2}\eta^{y_3+z_2}(1-\theta-\eta)^{y_5}$$

满条件分布为

$$\pi(\theta|y,z,\eta)\propto\theta^{z_1+y_2}[(1-\eta)-\theta]^{y_5}$$

$$\propto \left(\frac{\theta}{1-\eta}\right)^{z_1+y_2}\left(1-\frac{\theta}{1-\eta}\right)^{y_5}$$

$$\triangleq (1-\eta)Be(z_1+y_2+1, y_5+1) \tag{6.64}$$

$$\pi(\eta|y,z,\theta) = (1-\theta)Be(y_3+z_2+1, y_5+1) \tag{6.65}$$

$$\pi(z_1|y,\theta,\eta) = \pi(z_1|y_1,\theta) = b\left(y_1, \frac{2\theta}{2\theta+1}\right) \tag{6.66}$$

$$\pi(z_2|y,\theta,\eta) = \pi(z_2|y_4,\eta) = b\left(y_4, \frac{2\eta}{2\eta+3}\right) \tag{6.67}$$

(6.64)至(6.67)给出了 $\pi(x_i|x_{-i})$的表达式,由此可进行 Gibbs 抽样. Gelfand 和 Smith[17]对上述数据采用 Gibbs 抽样进行了计算,他们采用遍历均值来判断收敛性:在链中每隔 10 个数据取一个样本,每 20 个样本计算一次遍历均值,在经过 5000 次这样的计算后 Gibbs 抽样已经收敛,由此可得到下列估计

$$\mathrm{E}(\eta|y) = 0.123$$

$$\mathrm{E}(\theta|y) = 0.520$$

$$\mathrm{Var}(\eta|y) = 0.0065$$

$$\mathrm{Var}(\theta|y) = 0.018$$

另外,Gefland 和 Smith[16]还用高斯二次积分方法求出了 $\mathrm{E}(\eta|y)=0.123$和 $\mathrm{E}(\theta|y)=0.520$,二者一致.

例 6.22 再考虑例 6.14 和表 6.3 的数据,此处我们采用 Gibbs 抽样来进行分析.在例 6.14 中我们已指出 $\mu_i, i=1,\cdots,m$ 的满条件分布为正态分布,另外,由(6.52)式可以导出 μ,σ,τ 的满条件分布,分别为

$$(\mu|\sigma,\tau,Y,Z) \sim N(\hat{\mu}, \tau^2/m), \qquad \hat{\mu} = \frac{1}{m}\sum_{i=1}^{m}\mu_i$$

$$(\sigma^2|\mu,\tau,Y,Z) \sim I\chi^2(n,\hat{\sigma}^2), \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2$$

$$(\tau|\mu,\sigma,Y,Z) \sim I\chi^2(m-1,\hat{\tau}^2), \qquad \hat{\tau}^2 = \frac{1}{m-1}\sum_{i=1}^{m}(\mu_i-\mu)^2$$

其中 $I\chi^2(n,\delta)$ 表示自由度为 n,具有非中心参数 δ 的倒 χ^2 分布.据此可进行 Gibbs 抽样.

为便于判断收敛，我们平行地产生 10 条 Markov 链，在经过 100 次迭代后它们就已经收敛，从而后面的链可以合并起来使用. 在得到大量随机数后，我们可同时计算它的各种数字特征，比如，在收敛后再抽 2000 次，可算得各参数的后验分布的分位数，结果见表 6.5.

表 6.5　Gibbs 抽样算得的后验分位数

分位点	2.5%	25%	中位数	75%	97.5%
μ_1	58.9	60.6	61.3	62.1	63.5
μ_2	63.9	65.3	65.9	66.6	67.7
μ_3	66.0	67.1	67.8	68.5	69.5
μ_4	59.5	60.6	61.1	61.7	62.8
μ	56.9	62.2	63.9	65.5	73.4
σ	1.8	2.2	2.4	2.6	3.3
τ	2.1	3.6	4.9	7.6	26.6

§6.4.4　Metropolis-Hastings 方法

一类比 Gibbs 抽样更早出现，也更一般化的 MCMC 方法是 Metropolis-Hastings 方法. Metropolis 等人在 1953 年提出了一种构造转移核的方法[18]，Hastings 随后对之加以推广[19]，形成 Metropolis-Hastings 方法，其思路如下.

任意选择一个不可约转移概率 $q(\cdot,\cdot)$ 以及一个函数 $\alpha(\cdot,\cdot)$，$0<\alpha(\cdot,\cdot)\leqslant 1$，对任一组合 $(x,x')(x\neq x')$，定义

$$p(x,x')=q(x,x')\alpha(x,x'),\qquad x\neq x' \tag{6.68}$$

则 $p(x,x')$ 形成一个转移核.

此方法的实施比较直观：如果链在时刻 t 处于状态 x，即 $X^{(t)}=x$，则首先由 $q(\cdot\mid x)$ 产生一个潜在的转移 $x\to x'$，然后根据概率 $\alpha(x,x')$ 决定是否转移. 也就是说，在潜在转移点 x' 找到后，以概率 $\alpha(x,x')$ 接受 x' 作为链在下一时刻的状态值，而以概率 $1-\alpha(x,x')$ 拒绝转移到 x'，从而链在下一时刻仍处于状态 x. 于是，在有了 x' 后，我们可从 $[0,1]$ 上均匀分布抽一个随机数 u，则

$$X^{(t+1)} = \begin{cases} x', & u \leqslant \alpha(x, x') \\ x, & u > \alpha(x, x') \end{cases}$$

一般,分布 $q(\cdot|x)$ 称为建议分布(Proposal distribution).

因为我们的目标是使后验分布 $\pi(x)$ 成为平稳分布,因此,在有了 $q(\cdot,\cdot)$ 后,应选择一个 $\alpha(\cdot,\cdot)$ 使相应的 $p(x,x')$ 以 $\pi(x)$ 为其平稳分布. 一个最常用的选择是

$$\alpha(x, x') = \min\left\{1, \frac{\pi(x')q(x', x)}{\pi(x)q(x, x')}\right\} \tag{6.69}$$

此时,$p(x,x')$ 为

$$p(x, x') = \begin{cases} q(x, x'), & \pi(x')q(x', x) \geqslant \pi(x)q(x, x') \\ q(x', x)\dfrac{\pi(x')}{\pi(x)}, & \pi(x')q(x', x) < \pi(x)q(x, x') \end{cases} \tag{6.70}$$

我们先介绍 Metropolis-Hastings 算法的两个性质,然后讨论 $q(x,x')$ 的选择问题.

定理 6.11 由(6.70)式产生的 Markov 链是可逆的,即

$$\pi(x)p(x, x') = \pi(x')p(x', x) \tag{6.71}$$

且 $\pi(x)$ 是由(6.70)式确定的 Markov 链的平稳分布.

证明:若 $x=x'$,则(6.71)显然成立. 设 $x\neq x'$,则

$$\begin{aligned} \pi(x)p(x, x') &= \pi(x)q(x, x')\min\left\{1, \frac{\pi(x')q(x', x)}{\pi(x)q(x, x')}\right\} \\ &= \min\{\pi(x)q(x, x'), \pi(x')q(x', x)\} \\ &= \pi(x')q(x', x)\min\left\{\frac{\pi(x)q(x, x')}{\pi(x')q(x', x)}, 1\right\} \\ &= \pi(x')p(x', x) \end{aligned}$$

故(6.71)式成立,易有

$$\begin{aligned} \int \pi(x)p(x, x')\mathrm{d}x &= \int \pi(x')p(x', x)\mathrm{d}x \\ &= \pi(x')\cdot\int p(x', x)\mathrm{d}x \\ &= \pi(x') \end{aligned}$$

证毕.

由定理 6.11 可以看出，建议分布 $q(x,x')$ 可以取各种形式. 下面我们介绍一些常用的建议分布的选择方法.

(1) Metropolis 选择

Metropolis 曾考虑对称的建议分布[18]，即

$$q(x,x') = q(x',x), \qquad \forall\, x,x'$$

此时，$\alpha(x,x')$ 简化为

$$\alpha(x,x') = \min\left\{1,\frac{\pi(x')}{\pi(x)}\right\}$$

对称的建议分布是很常用的，比如，当 x 给定时，$q(x,x')$ 可取成正态分布，它以 x 为均值，而方差(协差阵)为常数(阵).

对称建议分布的一个特例是 $q(x,x')=q(|x-x'|)$，它称为随机移动 Metropolis 算法. 例如，

$$\pi(x) \propto \exp\{-x^2/2\}, \quad q(x,x') \propto \exp\{-(x'-x)^2/2\}$$

(2) 独立抽样

如果 $q(x,x')$ 与当前状态 x 无关，即 $q(x,x')=q(x')$，则由此建议分布导出的 Metropolis-Hastings 算法称为独立抽样. 此处，$\alpha(x,x')$ 变为

$$\alpha(x,x') = \min\left\{1,\frac{w(x')}{w(x)}\right\}$$

其中 $w(x)=\pi(x)/q(x)$.

一般，独立抽样的效果可能很好，也可能很不好[20]. 通常，要使独立抽样有好的效果，$q(x)$ 应接近 $\pi(x)$，比较安全的办法是使 $q(x)$ 的尾比 $\pi(x)$ 重.

(3) 单元素 Metropolis-Hastings 算法

同时产生整个 X 有时是困难的，而将 X 根据其分量进行逐个抽样则简单得多，这就要用到条件分布，特别是满条件分布.

考虑 $X_i|X_{-i}, i=1,\cdots,n$ 的条件分布，选择一个转移核 $q(x_i\to x'_i|x_{-i})$，固定 $X'_{-i}=X_{-i}=x_{-i}$ 不变，由 $q_i(x_i\to x'_i|x_{-i})$ 产生一个可能的 x'_i，然后以概率

$$\alpha_i(x_i\to x'_i|x_{-i}) = \min\left\{1,\frac{\pi(x')q_i(x'_i\to x_i|x_{-i})}{\pi(x)q_i(x_i\to x'_i|x_{-i})}\right\}$$

决定是否接受 x' 作为链的下一状态. 这就是单元素 Metropolis-Hastings 算法.

Gibbs 抽样是一种单元素 Metropolis Hastings 算法，它是在 Metropolis-hastings 方法中取 $q(x' \to x)$ 为 $\pi(x_i | x_{-i})$，容易验证，此时 $\alpha(x' \to x) = 1$.

在 Gibbs 抽样中，$\pi(x_i | x_{-i})$ 可能难于抽取，而 Metropolis 方法具有很大的灵活性，它可取 $q(\cdot, \cdot)$ 为易于抽取的分布. 有些文献研究了 Gibbs 抽样与 Metropolis 方法的结合问题，比如在 Gibbs 抽样中使用 Metropolis 方法来抽取随机数，Gilks，Best，和 Tan[11] 提出了一种 Gibbs 抽样中的调整筛选 Metropolis 抽样方法（Adaptive Rejection Metropolis Sampling），很有应用价值.

§6.4.5 应用

MCMC 算法使得 Bayes 方法中许多看起来困难的计算变得简单直观，下面介绍几个方面的应用.

（1）约束参数模型

对带约束的参数模型，一般的分析计算比较麻烦，而 Gibbs 抽样用于约束参数模型则非常简单. 在 θ_{-i} 给定下，由约束可给出 θ_i 的变动范围（约束参数空间的截面），它一般是一个区间（有时是区间的并），Gibbs 抽样只需要从限制在该区间内的一维分布抽样即可，它比直接进行带约束的高维积分简单得多.

比如，$\theta_1 \leqslant \theta_2 \leqslant \cdots \leqslant \theta_n$ 是一种常见的参数约束条件，若 θ 的后验分布是 $\pi(\theta | y)$，则满条件分布为

$$\pi(\theta_i | \theta_{-i}, y) \propto \pi(\theta | y) I_{\theta_{i-1} \leqslant \theta_i \leqslant \theta_{i+1}}$$

其中 $\theta_0 = -\infty, \theta_{n+1} = \infty$，由该分布抽样是直接的.

（2）变点问题

对多变点问题，Gibbs 抽样可将之简化. 记 $Y = (Y_1, \cdots, Y_m)$，若

$$Y_{r_{i-1}+1}, \cdots, Y_{r_i} \sim p_i(y | \theta_i), \quad i = 1, \cdots, k+1$$

其中 $\theta_1, \cdots, \theta_k, \theta_{k+1}$ 各不相等，$0 = r_0 < r_1 < \cdots < r_k < r_{k+1} = m$，则称 Y 有 k

个变点. 设 $r_1,\cdots,r_k,\theta_1,\cdots,\theta_{k+1}$ 未知，满条件分布是简单的. 记 $\boldsymbol{r}=(r_1,\cdots,r_k)$，不难看出

$$\pi(\theta_i|\theta_{-i},r,y)\propto\pi(\theta_i|y_{r_{i-1}+1},\cdots,y_{r_i})\tag{6.72}$$

$$\pi(r_i|r_{-i},\theta,y)\propto\pi(r_i|r_{i-1},r_{i+1},\theta,y)\tag{6.73}$$

(6.72)转化为无变点问题的后验分布，(6.73)转化为单变点问题且模型参数均已知. 看一个例子.

例 6.23 设 $Y_1,\cdots,Y_n$ 独立，且 $Y_i\sim P(\lambda t_i)$，$i=1,\cdots,k$，$Y_j\sim P(\theta t_j)$，$j=k+1,\cdots,n$，其中 $t_1,\cdots,t_n$ 已知，k,λ,θ 未知，这是单变点问题. 假设 k,λ,θ 的先验分布为：k 在 $\{1,\cdots,n\}$ 上均匀分布，$\lambda\sim Ga(a_1,b_1)$，$\theta\sim Ga(a_2,b_2)$，且 k,λ,θ 独立，a_1,a_2 已知，但 b_1,b_2 是未知超参数. 取第二层先验分布：$b_i\sim IG(c_i,d_i)$，$i=1,2$，并设 b_1,b_2 独立，其中 $IG(c,d)$表示倒 Gamma 分布，c_1,c_2,d_1,d_2 已知，则由 Bayes 公式，我们可写出诸满条件分布为

$$p(\lambda|\theta,k,b_1,b_2,Y)\sim Ga\left(a_1+\sum_{i=1}^{k}Y_i,\left(\sum_{i=1}^{k}t_i+b_1^{-1}\right)^{-1}\right)$$

$$p(\theta|\lambda,k,b_1,b_2,Y)\sim Ga\left(a_2+\sum_{i=k+1}^{n}Y_i,\left(\sum_{i=k+1}^{n}t_i+b_2^{-1}\right)^{-1}\right)$$

$$p(b_1|b_2,\lambda,\theta,k,Y)\sim IG(a_1+c_1,(\lambda+d_1^{-1})^{-1})$$

$$p(b_2|b_1,\lambda,\theta,k,Y)\sim IG(a_2+c_2,(\theta+d_2^{-1})^{-1})$$

$$p(k|\lambda,\theta,b_1,b_2,Y)\sim\frac{L(Y,\lambda,\theta,k)}{\sum_{k=1}^{n}L(Y,\lambda,\theta,k)}$$

其中

$$L(Y,\lambda,\theta,k)\propto p(Y|\lambda,\theta,k)\ (\text{可差一个与 } k \text{ 无关的因子})$$

$$\propto\lambda^{\sum_{i=1}^{k}Y_i}e^{-\lambda\sum_{i=1}^{k}t_i}\theta^{\sum_{i=k+1}^{n}Y_i}e^{-\theta\sum_{i=k+1}^{n}t_i}$$

$$\propto e^{(\theta-\lambda)\sum_{i=1}^{k}t_i}(\lambda/\theta)^{\sum_{i=1}^{k}Y_i}$$

我们在上面的推导中略去了具体的步骤，请读者自己完成. 有文献[10]中将上述满条件分布应用到一个具体的数据并进行迭代，给出了计算结果.

(3) 截尾数据和分组数据

明确起见，设 $y=(y_1,\cdots,y_s)$ 是观测值而 $y_{s+1},\cdots,y_m$ 只知其落在一个区间内，即 $v_i\leqslant y_i\leqslant u_i$，$i=s+1,\cdots,m$，记 $\boldsymbol{V}=(v_{s+1},\cdots,v_m)$，$\boldsymbol{U}=(u_{s+1},\cdots,u_m)$，则 θ 的后验分布为

$$\pi(\theta|\boldsymbol{V},\boldsymbol{U},y)\propto\pi(\theta)\prod_{i=1}^{s}p(y_i|\theta)\prod_{j=s+1}^{m}\int_{v_j}^{u_j}p(y_j|\theta)\mathrm{d}y_j \qquad (6.74)$$

其中 $\pi(\theta)$ 是 θ 的先验分布.

(6.74) 一般是难处理的，可引进添加数据 $z\triangleq(z_1,\cdots,z_{m-s})\triangleq(y_{s+1},\cdots,y_m)$，考虑 (θ,z) 的后验分布. 易见满条件分布为

$$\pi(\theta_i|\theta_{-i},z,\boldsymbol{V},\boldsymbol{U},y)\propto\pi(\theta|z,y) \qquad (6.75)$$

$$\pi(z_j|z_{-j},\theta,\boldsymbol{V},\boldsymbol{U},y)\propto\int_{v_{j+s}}^{u_{j+s}}p(x|\theta)\mathrm{d}x \qquad (6.76)$$

(6.75) 转化为无截尾数据的后验分布，而 (6.76) 则是由截断分布抽样.

更一般地，可考虑分组数据的处理. 设在 $[\tau_{i-1},\tau_i)$ 中有 r_i 个观测值（具体值未知），$i=1,\cdots,k+1$，$\tau_0=-\infty$，$\tau_{k+1}=\infty$，$\sum_{i=1}^{k+1}r_i=m$. 引进 $z=(z_1,\cdots,z_m)$，$z_1\leqslant z_2\leqslant\cdots\leqslant z_m$ 是未观测到的具体试验结果（缺损数据），则

$$\pi(\theta_i|\theta_{-i},z,\boldsymbol{\tau},\boldsymbol{r})\propto\pi(\theta|z) \qquad (6.77)$$

$$\pi(z_J|z_{-J},\theta,\boldsymbol{\tau},\boldsymbol{r})\propto p(z_{r_{j-1}+1}|\theta)\cdots p(z_{r_j}|\theta)I_{[\tau_{j-1}\leqslant z_{r_{j-1}+1}\leqslant\cdots\leqslant z_{r_j}<\tau_j]} \qquad (6.78)$$

其中 $J=\{r_{j-1}+1,\cdots,r_j\}$，$r_0=0$，$\boldsymbol{\tau}=(\tau_1,\cdots,\tau_k)$，$\boldsymbol{r}=(r_1,\cdots,r_k)$. 公式 (6.77) 是完全数据的后验分布，(6.78) 是截断分布的有序样本的抽样，可进一步采用 (1) 中的约束模型抽样方法.

MCMC 算法的应用范围是很广的，比如多层先验模型，广义线性模型，时间序列，多个分布的混合等等，有兴趣的读者可参阅有关文献[20].

参 考 文 献

1 肖云茹. 概率统计计算方法. 天津：南开大学出版社，1994

2 高惠璇. 统计计算. 北京：北京大学出版社，1995

3 Bulter J W. Machine sampling from given probabolity distributions. In: Symposium on Monte Carlo Methods, edited by Meyer M A, New York: Wiley, 1958

4 Rubinstein R Y. Simulation and Monte Carlo methods. New York: John Wiley-Sons, 1981

5 Box G E P, Muller M E. A note on the generation of random normal deviates. Ann Math Stat, 1958, 29: 610～611

6 Dempster A P, Laird N, Rubin D B. Maximam likelihood estimation from incomplete data via the EM algorithm (with discussion). J Roy Statist Soc B, 1977, 39: 1～38

7 Louis T A. Finding the observed information matrix when using the EM algorithm J Roy. Statist Soc B, 1982, 51: 127～138

8 Lansky D, Casella G. Improving the EM algorithm. Computing Science and Statistics: Proceedings of the twenty-Second Symposium on the interface. Washington DC: USA, 1990

9 Wei G C G, Tanner M A. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. J Amer Statist Assoc, 1990, 85: 699～704

10 Tanner M A. Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions. New York: Springer-Verlag, 1993

11 Gilks W R, Best N G, Tan K K C. Adaptive rejection Metropolis sampling within Gibbs sampling. Appl Statist, 1995, 44: 455～472

12 Gilks W R. Derivative-free adaptive rejection sampling for Gibbs sampling. in: Bayesian Statistics 4. Oxford: University Press, 1992, 641～649

13 Wu C F J. On the convergence properties of the EM algorithm. in: The Annals of Statistics, 1983, 11: 95～103

14 Meng X L, Rubin D B. Recent Extensions to the EM algorithm (with discussion). in: Bayesian Statistics 4. Oxford: University Press, 1992.307～320

15 Geman S, Geman D. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984, 6: 721～741

16 Gelfand A E, Smith A F M. Sampling based approaches to calculating marginal densities. J Amer Statist Assoc, 1990, 85: 398～409

17 Gelfand A E, Smith A F M. Gibbs sampling for marginal posterior expections. Communications in Statistics A, 1991, 20: 1747～1766

18 Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H, Teller E. Equations of state calculations by fast computing machines. J Chemical Phys, 1953, 21: 1087～1091

19 Hastings W K. Monte Carlo sampling methods using Markov Chains and their applications. Biometrika, 1970, 57: 72～89

20 Gilks W R, Richardson S, Spiegelhalter D J. Markov Chain Monte Carlo in practice. Chapman and Hall, 1996

习 题 六

6.1 设 $X\sim f(x)=3x^2, 0\leqslant x\leqslant 1$,试用逆变换法给出抽样步骤.

6.2 考虑一个分段表示的密度函数

$$f(x)=\sum_{i=1}^{n} f_i(x) I_{(x_{i-1}<x\leqslant x_i)}$$

其中 $a=x_0<x_1<\cdots<x_n=b$,a,b 可为无穷,试用逆变换法给出它的抽样方法.

6.3 设 $X_1,\cdots,X_n$ 独立同分布于某个 $F(x)$,$Y=\max\{X_1,\cdots,X_n\}$,试给出 Y 的抽样方法.可考虑一个具体的分布,比如 $F(x)=$

$1-e^{-x}$.

6.4　设 $X\sim f(x)=n\int_1^{\infty}y^{-n}e^{-xy}dy$,试用合成法给出一个抽样方案.

6.5　设 $X\sim F(x)=\sum_{i=1}^{\infty}p_i x^i$,$0\leqslant x\leqslant 1$,其中诸 $p_i\geqslant 0$ 已知,$\sum p_i=1$,试用合成法给出一个抽样方案.

6.6　证明:对任意的 $\alpha>0$,$\beta>0$ 和 $0\leqslant x\leqslant 1$,有

$$x^{\alpha-1}(1-x)^{\beta-1}\leqslant x^{\alpha-1}+(1-x)^{\beta-1}$$

利用此不等式,令

$$h(x)=\frac{\alpha\beta}{\alpha+\beta}[x^{\alpha-1}+(1-x)^{\beta-1}]$$

$$g(x)=[x^{\alpha-1}+(1-x)^{\beta-1}]^{-1}x^{\alpha-1}(1-x)^{\beta-1}$$

$$c=\frac{\alpha+\beta}{\alpha\beta}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

由此给出 Beta 分布的抽样方案.

6.7　对一些 n,比如 $n=3,10,20$,分别用(6.8)和(6.9)产生 n 个 $Exp(1)$ 随机数,比较它们所需的时间.

6.8　(1) 证明引理 6.4 的结论;

(2) 设 X_1,X_2 是相互独立的标准正态变量,令 $Y=X_1/X_2$,证明:Y 服从 Cauchy 分布,由此给出 Cauchy 分布的一个抽样方法.

6.9　设 X 服从 Laplace 分布(双指数分布),密度函数为

$$f(x)=\frac{1}{2}\theta e^{-|x-\mu|/\theta},\qquad \theta>0$$

试分别用逆变换法和合成法给出其抽样方案.

6.10　试用逆变换法产生来自极值分布的随机数.极值分布的密度函数为

$$f(x)=\frac{1}{\sigma}\exp\left\{-\frac{1}{\sigma}(x-\mu)-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\},\qquad \sigma>0$$

6.11　设 X 的分布密度函数为

$$f(x)=\begin{cases}\dfrac{x-2a}{(b-a)^2}, & 2a\leqslant x<a+b\\ \dfrac{2b-x}{(b-a)^2}, & a+b\leqslant x\leqslant 2b\\ 0, & 其它\end{cases}$$

试用逆变换法及合成法给出其抽样方案.

6.12 令

$$f(x)=\begin{cases}c_i x, & x_{i-1}\leqslant x<x_i\\ 0, & 其它\end{cases} \tag{6.79}$$

其中 $x_0=a\geqslant 0, x_n=b, c_i\geqslant 0, i=1,\cdots n$. 证明:若 $U\sim U(0,1)$,则 $X=[x_{i-1}^2+2(U-F_{i-1})/c_i]^{1/2}$ 的分布密度函数为(6.79)式.其中

$$F_i=\sum_{j=1}^{i}\int_{x_{j-1}}^{x_j}c_j x\mathrm{d}x, \qquad F_{i-1}<U\leqslant F_i$$

从而给出分布(6.79)的抽样方案.

6.13 设 X 是连续型随机变量,其密度函数及分布函数分别为 $f(x)$ 和 $F(x)$, $x\in(-\infty,\infty)$. 令

$$g(x)=\frac{f(x)}{F(b)-F(a)}, \quad x\in[a,b]$$

则 $g(x)$ 是由 $f(x)$ 定义的在 $[a,b]$ 上截断的分布. 试给出由 $g(x)$ 抽取随机数的方法.

6.14 设 X 的密度函数为

$$f(x)=\frac{3x^2}{2}, \qquad x\in(-1,1)$$

试用逆变换法,合成法和筛选法产生它的随机数,并比较这三种抽样方法的好坏.

6.15 设 U_1, U_2 来自 $U(0,1)$ 分布,令 $Y_1=U_1^{1/\alpha}, Y_2=U_2^{1/\beta}$,其中 $\alpha>0, \beta>0$. 证明,若 $Y_1+Y_2\leqslant 1$,则

$$X=\frac{Y_1}{Y_1+Y_2}$$

服从 Beta 分布 $Be(\alpha,\beta)$. 据此给出 Beta 分布的一个抽样方法.

6.16 考虑 χ^2 分布的一个抽样方法. 证明:若 $U_1,\cdots U_n$ 是来自

$U(0,1)$的随机数，令

$$X = -2\log \prod_{i=1}^{n} U_i \tag{6.80}$$

则 X 服从 $\chi^2(2n)$分布，从而(6.80)给出 $\chi^2(2n)$的一个抽样方法.

6.17　考虑积分 $I = \int_a^b f(x)g(x)\mathrm{d}x$，其中 $g(x)$是一个密度函数. 设 $X_1,\cdots,X_n$ 是来自 $U(a,b)$的随机数，则我们可由样本平均值法及关联抽样法给出 I 的估计为

$$\hat{I}_1 = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

$$\hat{I}_2 = \frac{1}{2n}\sum_{i=1}^{n}[f(X_i) + f(b + a - X_i)]$$

证明：若定理 6.8 的条件满足，则有 $\mathrm{Var}(\hat{I}_2)\leqslant\frac{1}{2}\mathrm{Var}(I_1)$.

6.18　试举例说明，如果分层不合理，也可能使抽样效果变差. 就是说，若在分配抽样次数时，贡献大的抽样次数反而少，则可能出现分层抽样给出的估计的方差比样本平均值法给出估计的方差还大.

6.19　试分别用随机投点法，样本平均值法，重要抽样法及分层抽样法近似计算 $\theta = \int_{-1}^{1} \mathrm{e}^x \mathrm{d}x$，并比较各种算法的精度.

6.20　记 $\theta = \int f(x)g(x)\mathrm{d}x$，$g(x)$ 是一个密度函数. 设 $h(x)$是另一密度函数，$X_1,\cdots,X_n$ 是来自 $h(x)$的样本，则 θ 的一个无偏估计是

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} f(X_i)\frac{g(X_i)}{h(X_i)}$$

证明：当 $h(x) = \dfrac{|f(x)|g(x)}{\int |f(x)|g(x)\mathrm{d}x}$时，$\mathrm{Var}(\hat{\theta}_h)$达最小，最小值为

$$\frac{1}{n}\{[\int |f(x)|g(x)\mathrm{d}x]^2 - \theta^2\}$$

6.21　设 X 服从 Cauchy 分布，其密度函数为

$$f(x) = [\pi(1 + x^2)]^{-1}$$

记 $\theta = P(X > 2)(=0.1476)$，考虑如下三个 θ 的估计：

(1) $\hat{\theta}_1=\frac{1}{n}\sum_{i=1}^{n}\varphi_1(x_i)$，其中

$\varphi_1(x)=\begin{cases}1, & \text{如果 } x>2,\\ 0, & \text{其它,}\end{cases}$ 而 $x_1,\cdots,x_n$ 是来自 Cauchy 分布的样本；

(2) $\hat{\theta}_2=\frac{1}{n}\sum_{i=1}^{n}\varphi_2(x_i)$，其中 $\varphi_2(x)=\begin{cases}1/2, & \text{如果 } |x|>2,\\ 0, & \text{其它,}\end{cases}$ 而 $x_1,\cdots,x_n$ 是来自 Cauchy 分布的样本；

(3) $\hat{\theta}_3=\frac{1}{n}\sum_{i=1}^{n}\varphi_3(x_i)$，其中 $\varphi_3(x)=\frac{1}{2}f(x)$，而 $x_1,\cdots,x_n$ 是来自 $U(0,1/2)$ 分布的样本.

试证明：$\hat{\theta}_j$，$j=1,2,3$ 都是 θ 的无偏估计，并比较它们的方差.

6.22 设 $X_i\sim N(\mu,\sigma^2/c_i)$，$i=1,\cdots,n$，诸 X_i 独立，其中 $c_1,\cdots,c_n$ 为独立同分布变量，其密度函数为 $f(c)$. 将 $c_1,\cdots,c_n$ 看作缺损数据，我们可用 EM 算法求 μ,σ^2 的估计.

(1) 导出该问题的 Q 函数；

(2) 记 μ,σ^2 的当前估计值为 $\mu^{(t)},\sigma^2_{(t)}$，证明 E 步简化为计算

$$w_i^{(t)}=\mathrm{E}(c_i|X_i,\mu^{(t)},\sigma^2_{(t)})$$

(3) 证明 M 步简化为计算

$$\mu^{(t+1)}=\frac{\sum_{i=1}^{n}w_i^{(t)}X_i}{\sum_{i=1}^{n}w_i^{(t)}}$$

$$\sigma^2_{(t+1)}=\frac{1}{n}\sum_{i=1}^{n}w_i^{(t)}(X_i-\mu^{(t)})^2$$

(4) 对一个具体的 c_i 的分布，譬如 χ^2 分布，给出 $w_i^{(t)}$.

6.23 令 $X_1,\cdots,X_n$ 为独立同分布变量，其密度函数为 $\lambda e^{-\lambda x}$. 设我们能够观测到的是 (Y_i,δ_i)，$Y_i=\min(X_i,c_i)$，$\delta_i=I_{(X_i\leqslant c_i)}$，$i=1,\cdots,n$. 取平坦先验分布(即 $\pi(\lambda)\propto 1$).

(1) 证明：若 λ 的当前估计值为 $\lambda^{(t)}$，则 Q 函数为

$$n\log\lambda-\lambda\sum_{i=1}^{n}\mathrm{E}(X_i|Y_i,\delta_i,\lambda^{(t)})$$

(2) 计算 $E(X_i|Y_i,\delta_i,\lambda^{(t)})$;

(3) 给出 M 步.

6.24　考虑 k 个分布的混合. 设 X_i 的分布密度函数为

$$f(x_i)=\sum_{j=1}^{k}p_j\varphi(x_i,\mu_j,\sigma),\qquad i=1,\cdots,n$$

其中 $\varphi(x,\mu_j,\sigma)=(\sqrt{2\pi}\sigma)^{-1}\exp\{-(x-\mu_j)^2/2\sigma^2\}$. p_j,μ_j,σ 都是未知的, $\sum_{j=1}^{k}p_j=1, 0<p_j<1$. 令 z_i 是一个 k 维示性向量, 表示 X_i 来自 k 个分量中的哪一个, 即 $z_i=(z_{i1},\cdots,z_{ik})$, 诸 z_{ij} 非 0 即 1, 且 $\sum_{j=1}^{k}z_{ij}=1$. 于是可给出添加似然函数为

$$\prod_{j=1}^{k}p_j^{z_{ij}}(\varphi(x_i,\mu_j,\sigma))^{z_{ij}}$$

(1)导出 Q 函数, 给出 E 步和 M 步;

(2)将表 6.6 中数据用上述模型拟合, 并用 EM 算法估计诸参数(取 $k=3$);

(3)计算估计的标准差.

表 6.6　数　据

9172	9350	9483	9558	9775	10227
10406	16084	16170	18419	18552	18600
18927	19052	19070	19330	19343	19349
19440	19473	19529	19541	19547	19663
19846	19856	19863	19914	19918	19973
19989	20166	20175	20179	20196	20215
20221	20415	20629	20795	20821	20846
20875	20986	21137	21492	21701	21814
21921	21960	22185	22209	22242	22249
22314	22374	22495	22746	22747	22888
22914	23206	23241	23263	23484	23538
23542	23666	23706	23711	24129	24285
24289	24366	24717	24990	25633	26960
26995	32065	32789	34279		

6.25　设 $y_i\sim b(m_i,\theta_i), i=1,\cdots,n$, 诸 y_i 独立. 假设

$$\theta_i=\frac{1}{1+\exp\{-(\alpha+\beta x_i)\}}$$

取平坦先验分布 $\pi(\alpha,\beta)\propto 1$.

(1) 导出 α、β 的后验分布;

(2) 设计一个单元素 Metropolis-Hastings 算法,导出其接收概率;

(3) 考虑能否对之加以改进.

6.26 考虑例 6.13 中的模型和数据,取平坦先验分布.

(1) 采用 Gibbs 抽样对它进行分析,给出 θ 的后验均值估计,并将迭代过程中的 $\theta^{(i)}$ 关于 i 作图;

(2) 同时从几个(譬如 10 个)不同的初始点出发进行 Gibbs 抽样,计算各自给出的 θ 的估计并作图.它们在经过一段迭代后是否趋于等同?这是一种常用的判断 Gibbs 抽样收敛性的方法;

(3) 将例 6.13 中的数据(125,18,20,34)改为(14,0,1,5),重复(1),(2).

6.27 仍考虑例 6.13 中的模型和数据,取平坦先验分布.试用 Metropolis-Hastings 算法进行分析,分别采用如下一些建议分布:$U(0,1)$;$N(\theta^{(i)},0.01)$,($\theta^{(i)}$ 为当前估计值,下同);$N(\theta^{(i)},0.0001)$;$N(0.4,0.01)$.给出估计值并作图.